国家图书馆民国报纸数字资源建设项目

国家图书馆　数字资源部

民国报纸数字资源建设

序号	项目年	报纸名称	记录标识号	出版日期起止号	期数	版数
1	2014-2015	大刚报	00N001037	19451109-19511231	2097	10173
2	2014-2015	阵中日报	00N001428	19390121-19490414	3426	7972
3	2014-2015	解放日报	00N002542	19410516-19470327	2130	7992
4	2014-2015	贵州商报	00N001674	19410430-19491108	969	3878
5	2014-2015	甘肃民国日报	00N001024	19330609-19490812	4933	20498
6	2014-2015	中央日报-贵阳	00N001745	19381201-19481229	3214	14597
7	2014-2015	南宁民国日报	00N001453	19310701- 19441027	2567	17507
8	2014-2015	新天津	00N000470	19241224-19440430	4689	53946
9	2014-2015	西北文化日报	00N001021	19310521-19490221	5865	24939
10	2014-2015	革命日报	00N001755	19380426-19491110	3578	14530
11	2014-2015	工商日报	00N000474	19370201-19430131	2036	5282
12	2014-2015	福建日报	00N001445	19390401-19481201	1962	5124
13	2014-2015	冀中导报	00N001724	19460812-19481231	841	3204
14	2014-2015	益世报	00N000203	19170128-19481225	8847	72488
15	2014-2015	新华日报	00N000846	19380111-19470228	3231	12447
16	2014-2015	华北日报	00N001043	19290101-19490131	3934	34772
17	2014-2015	中央日报-昆明	00N001720	19390515-19491209	3724	17580
18	2014-2015	青岛时报	00N001721	19320501-19480625	1891	21793
19	2014-2015	中央日报-福建	00N001044	19410428-19490816	1405	7666
20	2014-2015	山东民国日报	00N000638	19290901-19460630	784	6164
21	2014-2015	国风日报	00N001694	19371010-19481031	2010	6234

民国报纸数字资源建设

- 文献前整理
- 数字化加工
- 数据验收

文献前整理

文献前整理

报种层面

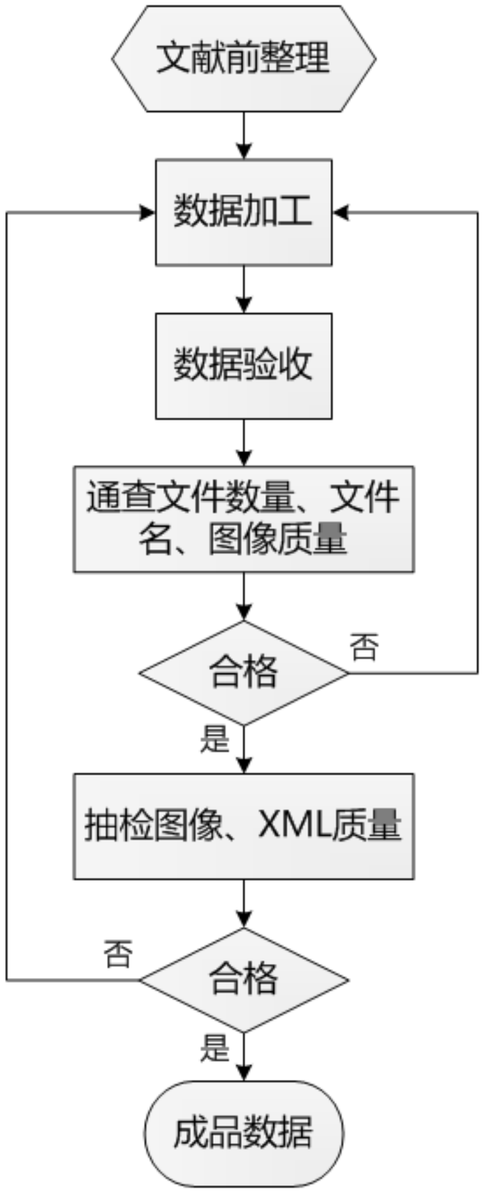
- 记录标识号；
- 报名是否前后变化；
- 版式特点
- ……

特种刊

- 增刊、特刊、副刊、号外等；
- 特种刊出版频率；
- 特种刊版面特点；
- 联合版情况；
- ……

印刷问题

- 卷期印错；
- 出版日期印错；
- 版次印错；
- 残缺等；
- ……



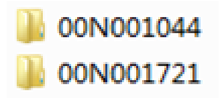
民国报纸数字资源建设

文献前整理

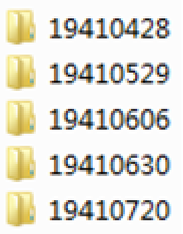
图像分类分包

缩微胶卷扫描后的民国报纸图像一般按流水号命名，如00001、00002……，一次扫描出的多卷图像存储在一个文件夹内，不利于同种报纸的存储和出版日期区分。因此，需按报种和按出版日期对民国报纸进行整理。

种号



出版日期



图像原始问题登记

在分类和分期的过程中，记录图像残缺、图像重复、疑似缺版和错版等情况。

包名	问题	卷期	页码
DG19470917	原9月17日第4版印刷成9月16日	大刚报 17：6	73
DG19471005	原10月5日第4版印刷成10月4日	大刚报 17：6	146
DG19471229	第3版版次印刷为第5版	大刚报 17：6	485
DG19480827	原8月27日第4版印刷成8月26日	大刚报 17：8	238
DG19481010	第2版版次模糊看不清	大刚报 17：8	410、411、412、413
DG19490516	缺少2、3版，只有1、4版	大刚报 17：10	
DG19491017	第1版版次印刷为第5版	大刚报 17：10	268
DG19491130	缺少1、2、3、4版，只有5、6版	大刚报 17：10	
DG19510901	第4版没有印刷日期	大刚报 17：17	76
DG19510902	原9月2日第4版印刷成9月1日	大刚报 17：17	82

原始问题核对

对记录的问题进行原始胶卷阅览核对，避免因误操作导致错误。

民国报纸数字资源建设

- 文献前整理
- 数字化加工
- 数据验收

数字化加工

- 依据整理结果，对整理好的民国报纸进行元数据加工、图像扫描、图像处理、OCR文字识别、及生成最终的图像、文本及双层PDF文件等。

基本要求

- 元数据：报纸的著录单位为一种报纸，涉及著录对象、著录内容、信息源等；
- TIFF图像数字化标准和命名规则；
- 双层PDF文件标准；
- XML文件标准；
- 数据库及说明文件等。

元数据的著录对象

字段号	字段说明	必备性	可重复性
001	著录数字馆藏记录标识号。001 字段的取值规则为 10 位流水号。由数字资源部根据项目合同分配号段。	M	NR
005	自动生成，著录数字资源记录的最后处理日期和时间。	O	NR
035	著录其他系统控制号。	O	R
100	除入档时间，其他可套用实体资源记录。	M	NR
101	套用实体资源记录。	M	NR
102	套用实体资源记录。	O	NR
106	删除实体资源记录中的形态特征编码数据字段。	不适用	不适用
121	删除实体资源记录中的形态特征编码数据字段。	不适用	不适用
126	删除实体资源记录中的形态特征编码数据字段。	不适用	不适用
130	删除实体资源记录中的形态特征编码数据字段。	不适用	不适用
141	删除实体资源记录中的形态特征编码数据字段。	不适用	不适用
194	删除实体资源记录中的形态特征编码数据字段。	不适用	不适用
135	著录有关数字资源的编码数据，如色别、声音、复制来源等。如有多种文件，重复该字段。	M	R
181	统一赋值：*181#0@ai4@bxxxe##*	O	R
182	统一赋值：*182#0@ab*	O	R
200	去除@b 一般资料标识的内容。 其他套用实体资源记录。	M	NR
203	统一赋值：*203##@a 文本@b 视觉@b 报纸@c 电子*	O	R
205	套用实体资源记录。	O	R
207	套用实体资源记录。	A	NR
210	套用实体资源记录。	O	NR
215	套用实体资源记录。	O	R
225	套用实体资源记录。	O	R
3XX	324 字段根据项目合同著录： *复制自：馆藏缩微胶卷*	324 字段为 M，其他 3XX	322、324、345 字段为 NR，其他

	或 *复制自：馆藏缩微平片* 其他 3XX 字段套用实体资源记录。	字段为 O	3XX 字段为 R
4XX	使用 452 字段著录馆藏实体资源元数据记录标识号（即 ALEPH01 库 001 字段。如原实体馆藏记录未入 ALEPH 系统，使用 452 字段关联其系统控制号。），将受编的数字资源与相关的实体资源建立连接。如： 452#0@0(NLC01)003342444 452#0@0z-024041 其他套用实体资源记录。	O	R
5XX	套用实体资源记录。	O	R
6XX	针对编目历史规则变更引起的分类号版本不统一进行分类号调整，消除分类号对文献进行汇集的质量影响。 其他套用实体资源记录。	O	R
7XX	套用实体资源记录。	A	700、710、720 字段为 NR，其他 7XX 字段为 R
801	著录数字馆藏的记录来源。 统一赋值为：801#0@aCN@bNLC	M	R

特殊情况的处理

1、**报名变更**：MARC记录200$a字段内容作为XML的报纸“题名”。另一名称著录为“题名备注”。

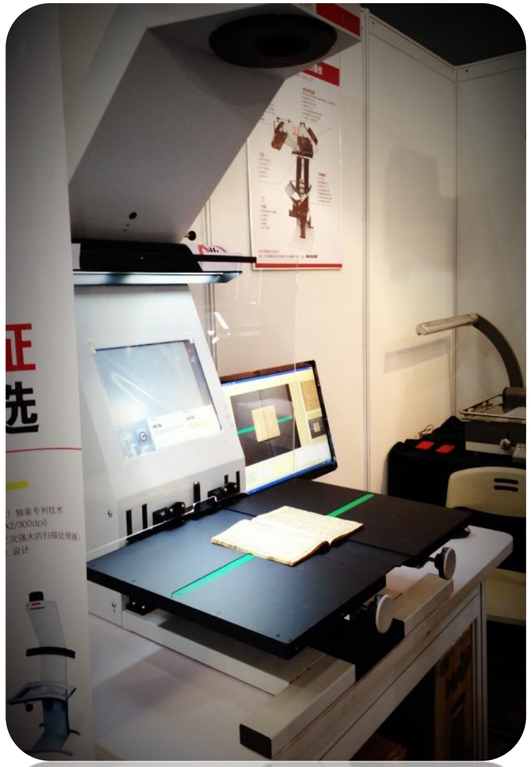
例：《革命日报》于1935年创刊，1940年1月1日更名为《贵州日报》，著录方式如图所示：

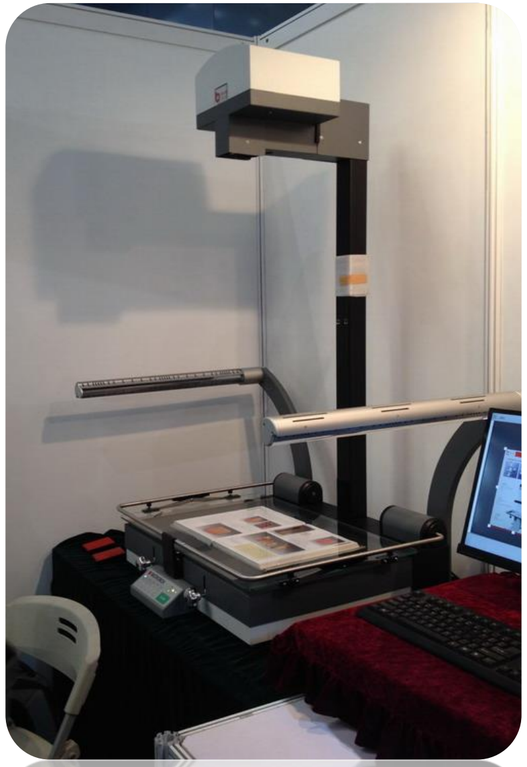

```
<?xml version="1.0" encoding="utf-8" ?>
<root>
    <报纸元数据>
        <记录标识号>00N001755</记录标识号>
        <题名>贵州日报</题名>
        <出版日期>19400101</出版日期>
        <卷期>1813</卷期>
        <版次>1</版次>
        <整版PDF链接>001.pdf</整版PDF链接>
        <备注>
            <题名备注>革命日报</题名备注>
        </备注>
    </报纸元数据>
```

民国报纸数字资源建设

数字化加工

高清拍照扫描仪





高速扫描仪

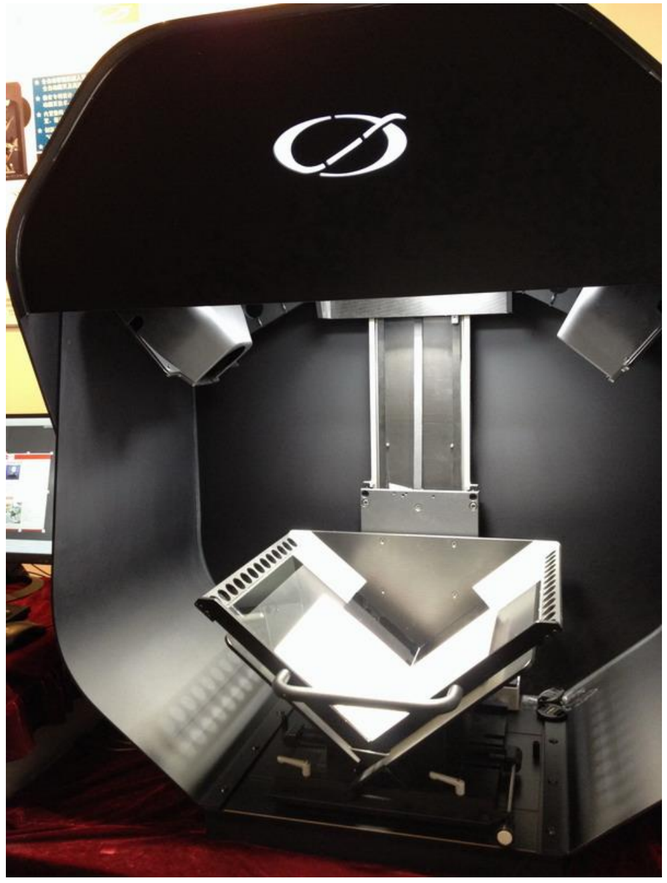
零边距扫描仪

滚筒扫描仪

民国报纸数字资源建设

数字化加工

缩微胶卷高速扫描仪



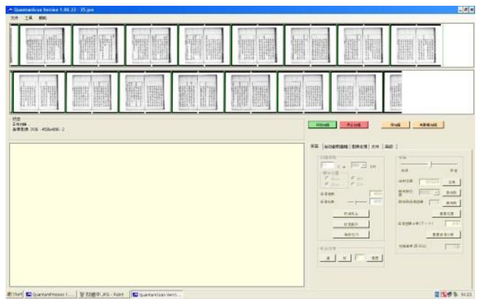
V型扫描仪

民国报纸数字资源建设

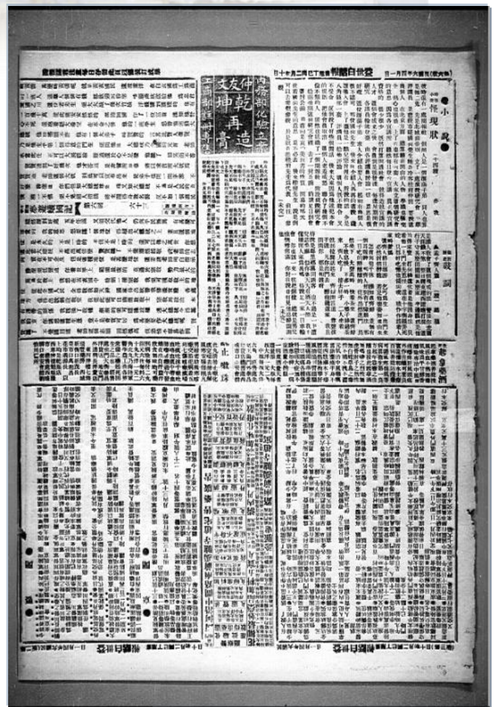
数字化加工

缩微技术

扫描设备



MEKEL MACH V 高速扫描仪
及相应处理软件



获得最终的原始数字图像，TIF格式、分辨率300dpi、位深8位等。

民国报纸数字资源建设

数字化加工

文献扫描

- 黑白页和灰度页用灰度方式扫描。

黑白页和灰度页用灰度方式扫描

色彩位深：8 位

分辨率：300 dpi

档案典藏级格式：TIFF 不压缩

发布服务级格式：XML、双层 PDF

- 彩色页用彩色方式扫描。

彩色页用彩色方式扫描

色彩位深：24 位

分辨率：300 dpi

档案典藏级格式：TIFF 不压缩

发布服务级格式：XML、双层 PDF

图像质量要求

数字化环境注意防护光源，避免透光或反射光的影响

1、图像清晰，亮度适中，由于原件造成的亮度不适可算合格；

2、扫描后的图像需进行去污、纠偏、去图像黑边等处理；

3、扫描后的图像要求真实反映原件，按版次顺序由小到大，符合阅读习惯，不能有缺版、错版、数据内容缺失等现象，原件问题除外；

4、图像综合错误率不超过1‰。

民国报纸数字资源建设

数字化加工

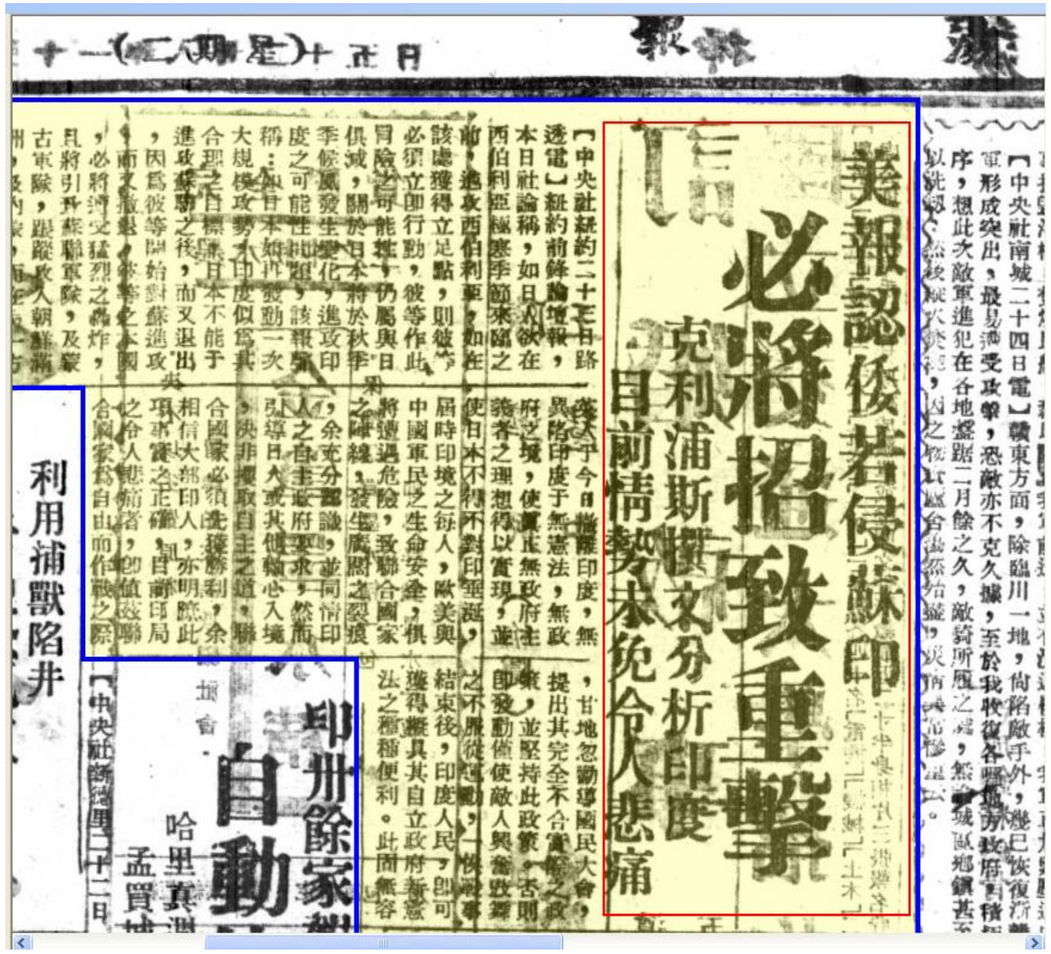
图片反射光源

无反射情况

民国报纸数字资源建设

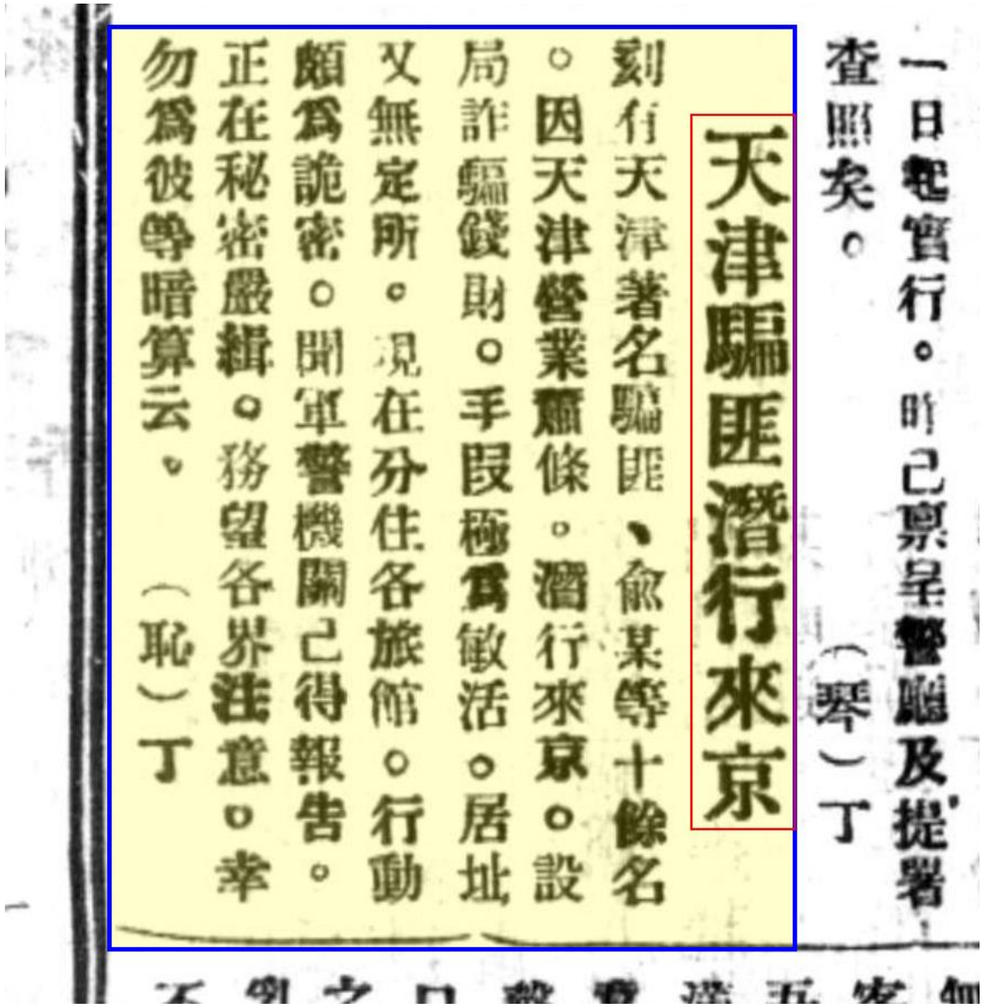
数字化加工

透字问题



美報認後若侵蘇印
必將招致重擊
克利浦斯提文分析印度
目前情勢未免令人悲痛
利用捕獸陷井
自動
哈里真
孟買

无透字状态



天津騙匪潛行來京

刻有天津著名騙匪、俞某等十餘名。因天津營業廳條。潛行來京。設局詐騙錢財。手段極爲敏活。居址又無定所。現在分住各旅館。行動頗爲詭密。開軍警機關已得報告。正在秘密嚴緝。務望各界注意。幸勿爲彼等暗算云。（恥）丁

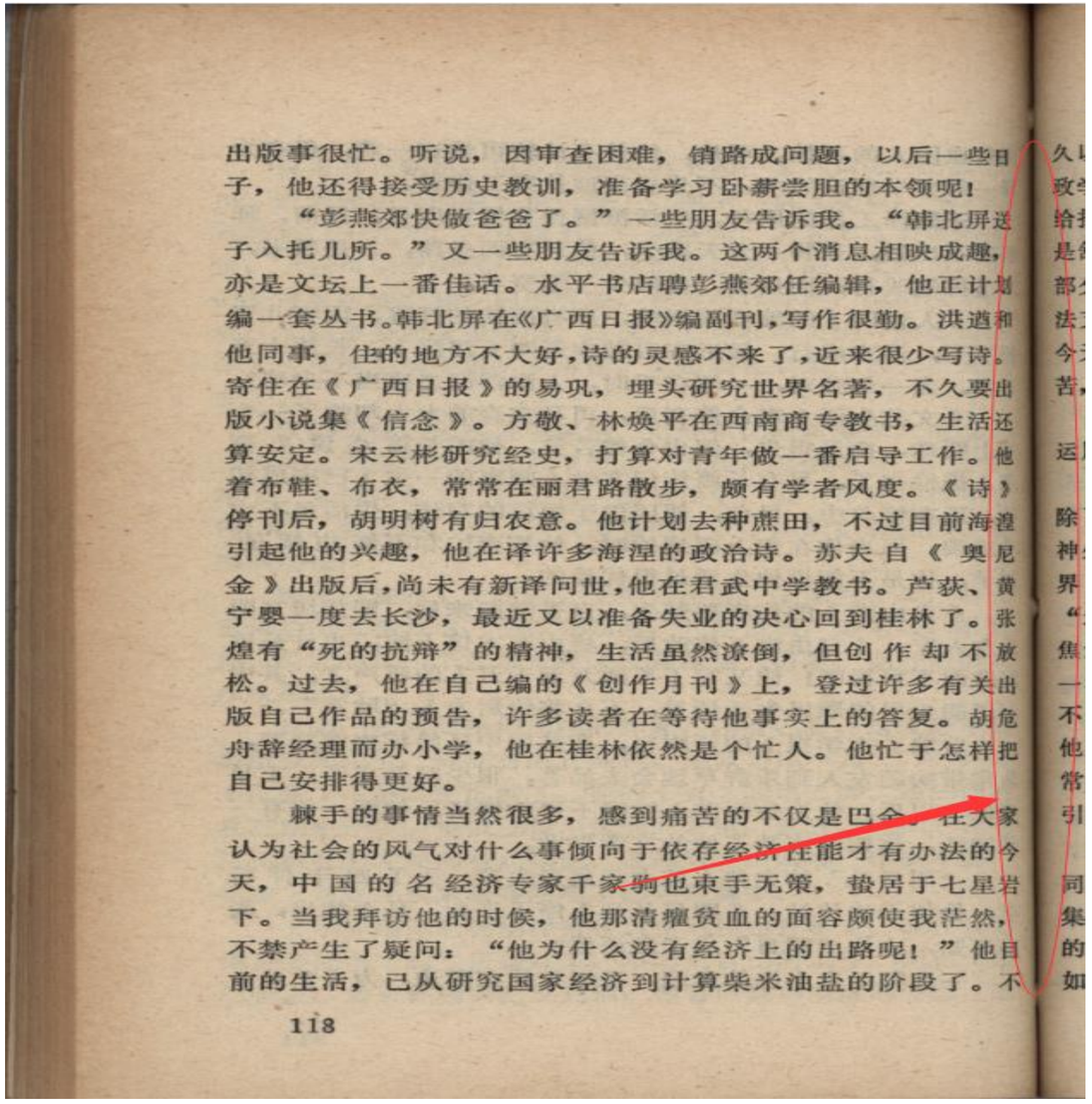
一日屯實行。昨已票呈警廳及提署查照矣。（琴）丁

民国报纸数字资源建设

数字化加工

以书为例子：文字偏斜，变形，边距不当

报纸一般会**拆装**扫描



出版事很忙。听说，因审查困难，销路成问题，以后一些日子，他还得接受历史教训，准备学习卧薪尝胆的本领呢！

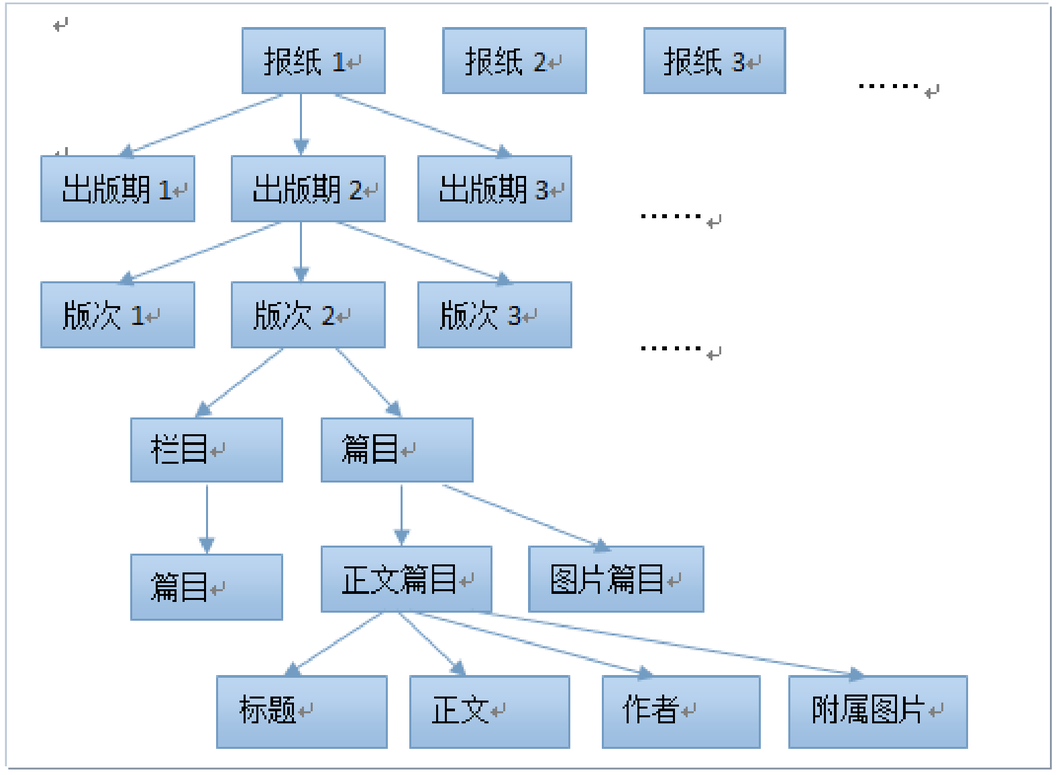
“彭燕郊快做爸爸了。”一些朋友告诉我。“韩北屏送子入托儿所。”又一些朋友告诉我。这两个消息相映成趣，亦是文坛上一番佳话。水平书店聘彭燕郊任编辑，他正计划编一套丛书。韩北屏在《广西日报》编副刊，写作很勤。洪遒和他同事，住的地方不大好，诗的灵感不来了，近来很少写诗。寄住在《广西日报》的易巩，埋头研究世界名著，不久要出版小说集《信念》。方敬、林焕平在西南商专教书，生活还算安定。宋云彬研究经史，打算对青年做一番启导工作。他着布鞋、布衣，常常在丽君路散步，颇有学者风度。《诗》停刊后，胡明树有归农意。他计划去种蔗田，不过目前海涅引起他的兴趣，他在译许多海涅的政治诗。苏夫自《奥尼金》出版后，尚未有新译问世，他在君武中学教书。芦荻、黄宁婴一度去长沙，最近又以准备失业的决心回到桂林了。张煌有“死的抗辩”的精神，生活虽然潦倒，但创作却不放松。过去，他在自己编的《创作月刊》上，登过许多有关出版自己作品的预告，许多读者在等待他事实上的答复。胡危舟辞经理而办小学，他在桂林依然是个忙人。他忙于怎样把自己安排得更好。

棘手的事情当然很多，感到痛苦的不仅是巴金。在大家认为社会的风气对什么事倾向于依存经济性能才有办法的今天，中国的名经济专家千家驹也束手无策，蛰居于七星岩下。当我拜访他的时候，他那清癯贫血的面容颇使我茫然，不禁产生了疑问：“他为什么没有经济上的出路呢！”他目前的生活，已从研究国家经济到计算柴米油盐的阶段了。不

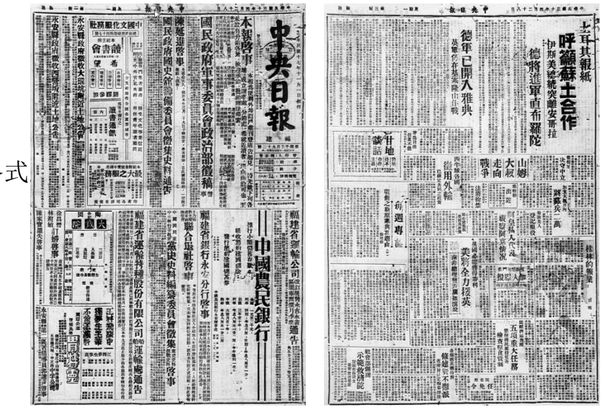
民国报纸数字资源建设

数字化加工

存储与命名



JPG格式



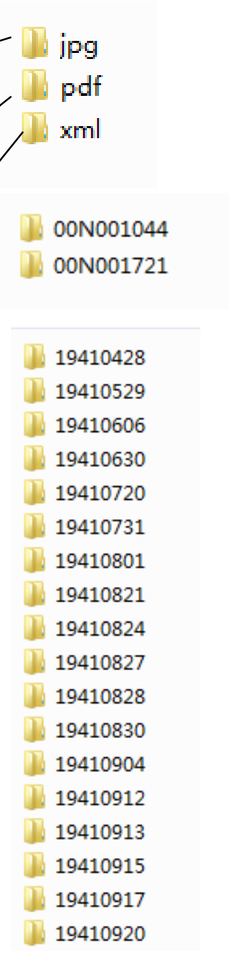
PDF格式

XML格式

格式

种

期



民国报纸数字资源建设

数字化加工

命名规则

某民国报纸记录标识号（种号）为00N001044，出版日期为1939年5月15日，报纸常规版面为4版，第4版分两拍拍摄，有增刊独立版面2版，则民国报纸相应的图像扫描文件命名和存储如下：

\ 00N001044 \19390515\001.tif 第1版

002.tif 第2版

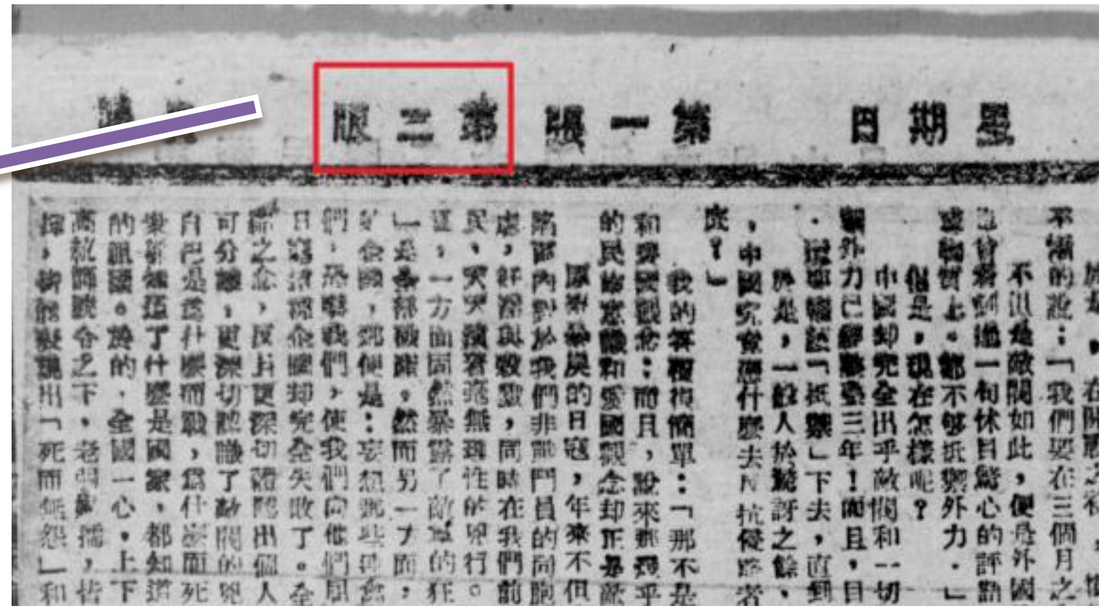
003.tif

004_1.tif

004_2.tif

Z01.tif 增刊第1版

Z02.tif



特殊情况的处理

2、特殊版面：

例：《益世报》副刊《女子周刊》一张上有两版，通版处理。

特殊情况的处理

2、特殊版面：

《新天津》的1939年2月22日出版的第十一版和第十二版报纸没有明显的版边界线，很多篇目内容贯穿两版。

特殊情况的处理

例：《益世报》副刊《益世白话报》，一张上有四小版，且部分内容文字方向不符合阅读习惯。

特殊情况的处理

例：《大刚报》1946年3月13日第四版的报纸，存在一个刀把型的规则空白区，其形成原因推测为剪报所致，且无法判断该区域覆盖几篇篇目文章

特殊情况的处理

例：《益世报》号外，纸张比常规版面略小。


益世報
號外
中華民國二十年十月十一日

國聯對日本越軌行動謀制止
行政院會決提前後日即開會
美對中日糾紛將正式發言反對軍事行動
德外部命駐瀋領返瀋調查輿論對我同情

國聯理事會對中日兩國通牒

美政府對日將再提嚴重覺書

日軍部所持態度仍異常强硬

民国报纸数字资源建设

数字化加工

特殊情况的处理

- 1946年2月2日至2月6日春节共5天，《大刚报》、《华中日报》、《新湖北日报》、《武汉日报》、《和平日报》共同出版了**《汉口各报联合版》**报纸。
- 广告的处理。
- 无标题需自拟。

民国报纸数字资源建设

数字化加工

成品数据

TIFF图像


貴州商報

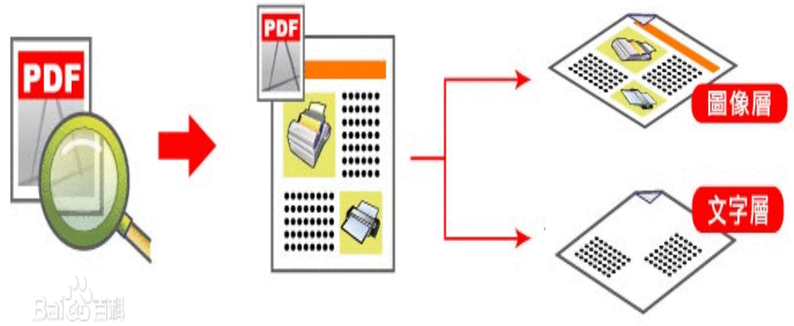
TIFF图像单版

XML文件

```
<?xml version="1.0" encoding="utf-8" ?>
<root>
	<报纸元数据>
		<记录标识号>00N000203</记录标识号>
		<题名>益世报</题名>
		<出版日期>19230611</出版日期>
		<卷期>2598</卷期>
		<版次>7Y</版次>
		<整版PDF链接>007.pdf</整版PDF链接>
	</报纸元数据>
	<版元数据>
		<正文组>
			<正文 篇目号="001">
				<栏目>小言</栏目>
				<引题 />
				<标题>小學教員與軍警</标题>
				<副题 />
				<标题坐标>3121,437,3174,437,3174,794,3121,794</标题坐标>
				<篇目坐标>2180,334,3174,334,3174,1374,2180,1374</篇目坐标>
				<转版信息 />
				<作者 />
				<附图组 />
			</正文>
			<正文 篇目号="002">
				<栏目>京畿寫實</栏目>
				<引题 />
				<标题>殺夫索燕骨驗屍</标题>
				<副题 />
				<标题坐标>1986,395,2036,395,2036,752,1986,752</标题坐标>
				<篇目坐标>404,286,2036,286,2036,1446,404,1446</篇目坐标>
				<转版信息 />
				<作者>止</作者>
				<附图组 />
			</正文>
			<正文 篇目号="003">
				<栏目>京畿寫實</栏目>
				<引题 />
				<标题>穆家圈二次發現炸彈</标题>
				<副题 />
				<标题坐标>344,395,386,395,386,857,344,857</标题坐标>
				<篇目坐标>296,276,386,276,386,1430,296,1430;2627,1446,3326,1446,3326,2618,2627,2618</篇目坐标>
				<转版信息 />
				<作者>止</作者>
				<附图组 />
			</正文>
			<正文 篇目号="004">
				<栏目>京畿寫實</栏目>
				<引题 />
				<标题>爭家產演唱劈棺</标题>
				<副题 />
				<标题坐标>2554,1568,2616,1568,2616,1974,2554,1974</标题坐标>
				<篇目坐标>1931,1463,2616,1463,2616,2610,1931,2610</篇目坐标>
				<转版信息 />
				<作者>言</作者>
				<附图组 />
			</正文>
```

单版XML

双层PDF



单版双层PDF

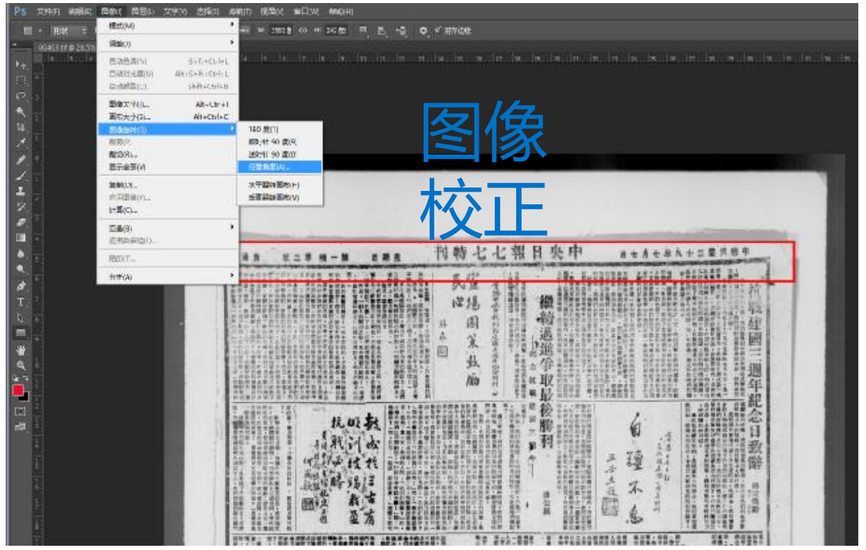
图像处理

原始TIF图像特点

原始TIF图像存在尺寸大小不一，有黑边，图像倾斜，版心不居中等质量问题，需进行图像处理得到长期保存级TIF图像。

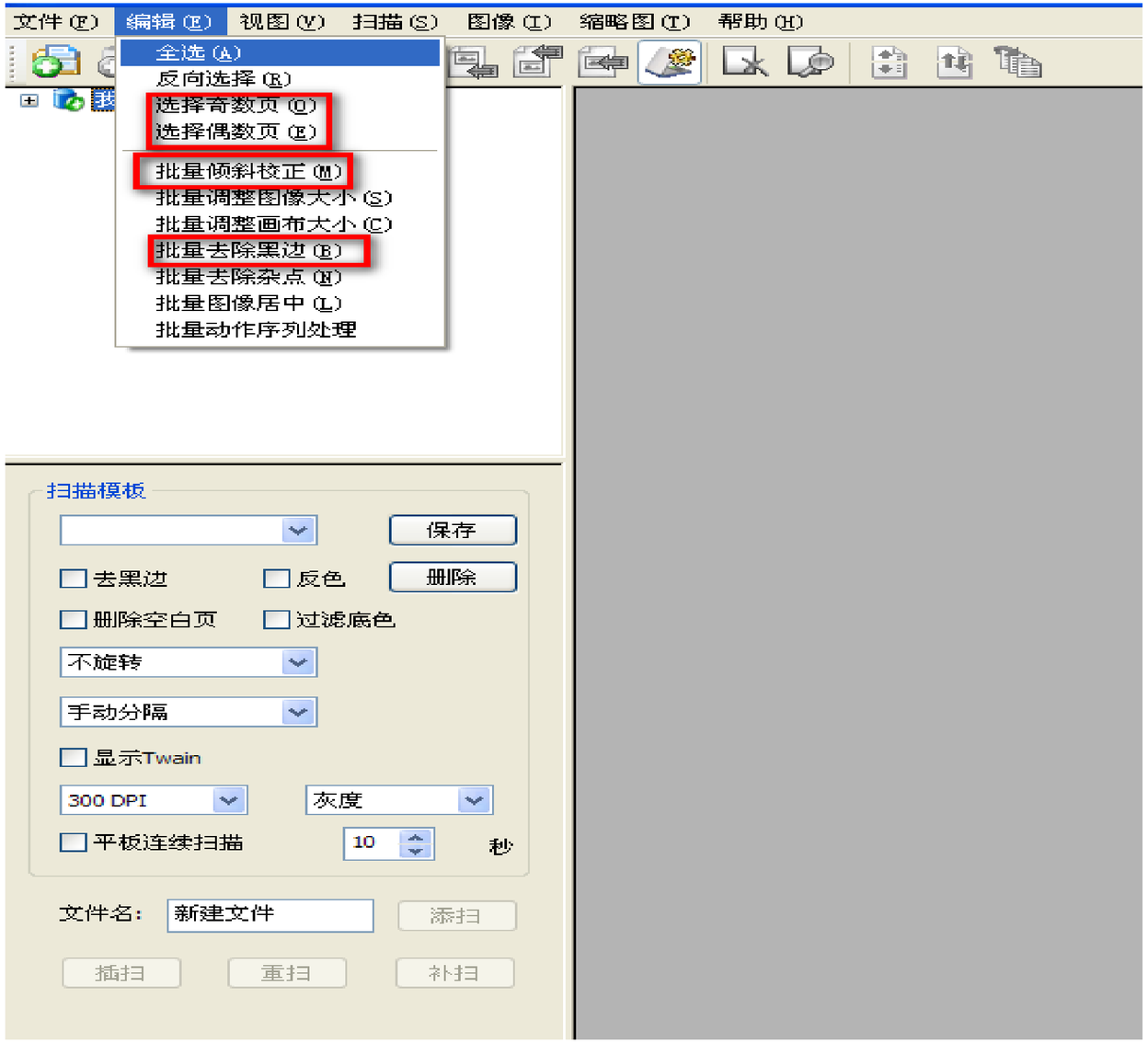
需进行的图像处理操作

1、图像校正
2、裁剪，去掉多余黑边
3、填充画布大小（使文字上下左右居中）
4、批量修改尺寸
5、压缩（300dpi）
6、检查质量



图像处理

自动检查整理图像的DPI与图像模式，批量进行倾斜校正，批量去除黑边。



民国报纸数字资源建设

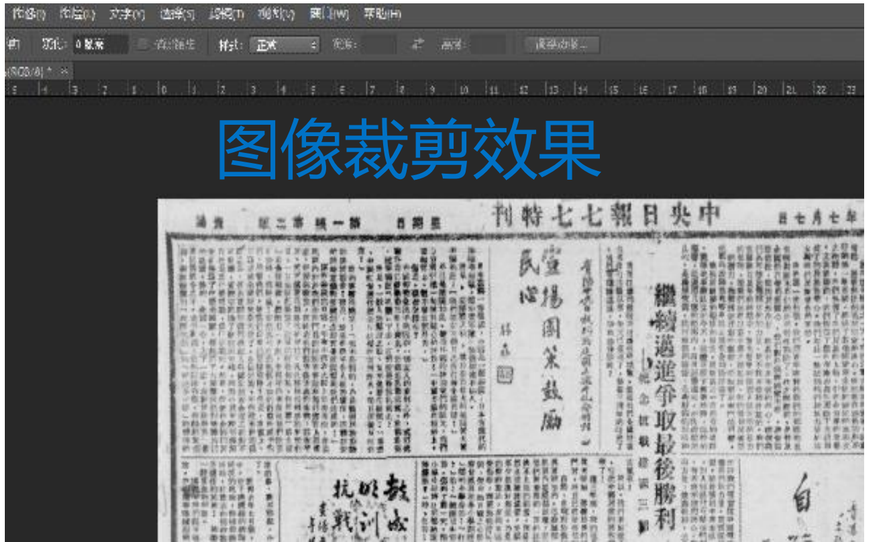
数字化加工

图像处理

图像裁剪

画布填充

图像裁剪效果



画布填充效果

民国报纸数字资源建设

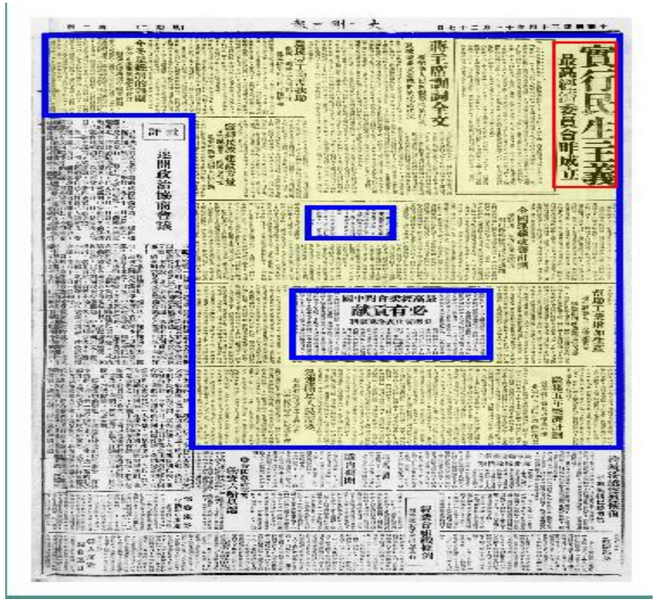
数字化加工

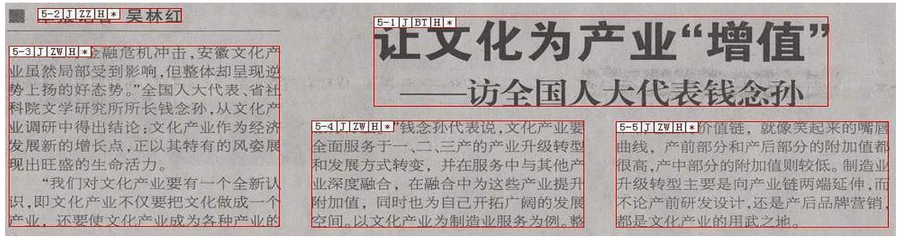
版面分析

版式分析

程序自动
程序判断划框区域是横排文本区、竖排文本区、表格区还是图像区，给以标识

人工手动
对程序结果进行干预或因版面复杂，最初即选择手动分析，获得标题、正文、作者等位置



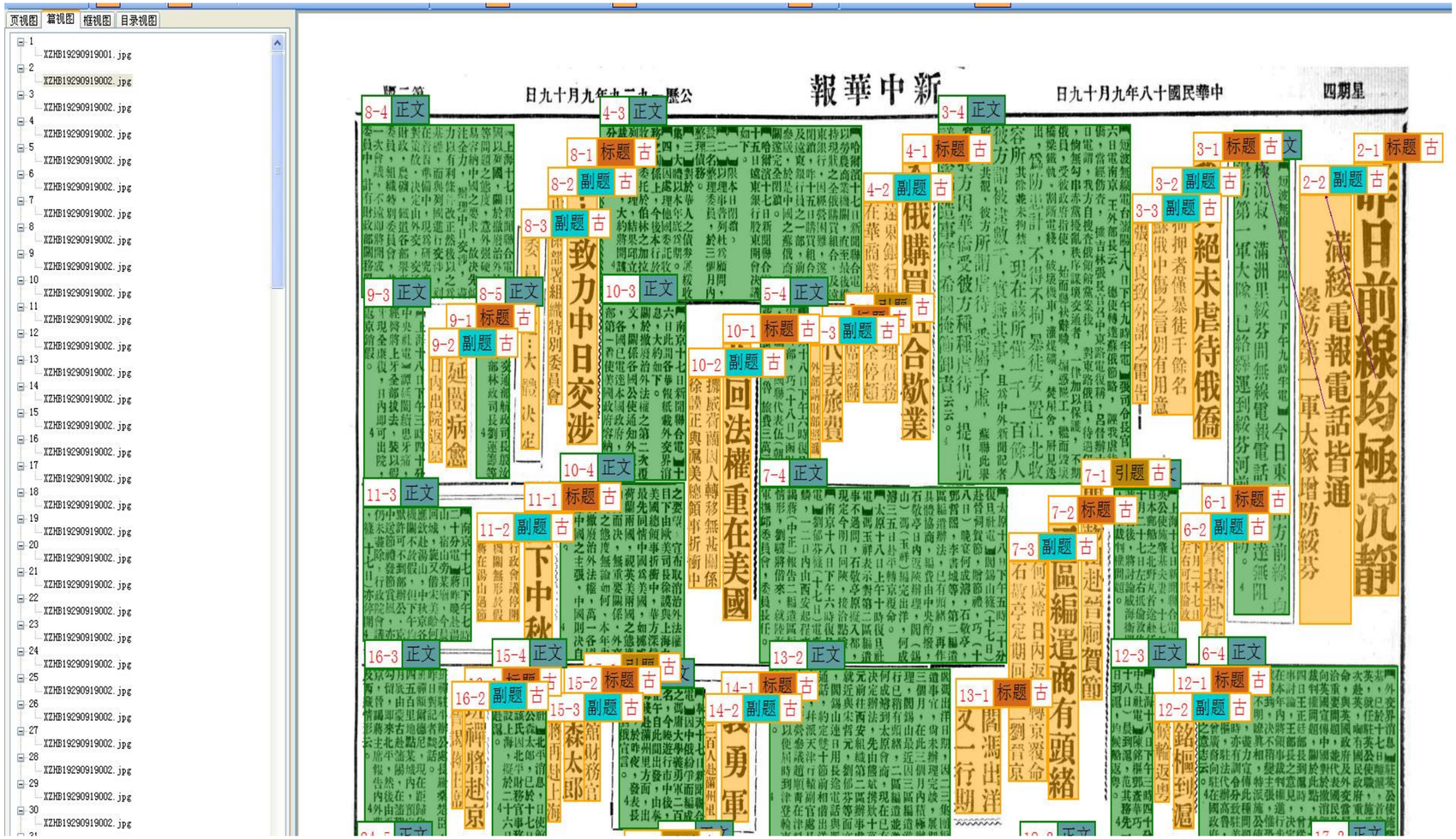


民国报纸数字资源建设

数字化加工

版面分析

对报纸上各栏目、篇目、篇目标题、作者等进行标识



切分校对

为能使后期PDF文字层与图像层对位准确，增加了此工序，做到一字一框，精确对位

字符聚类校对

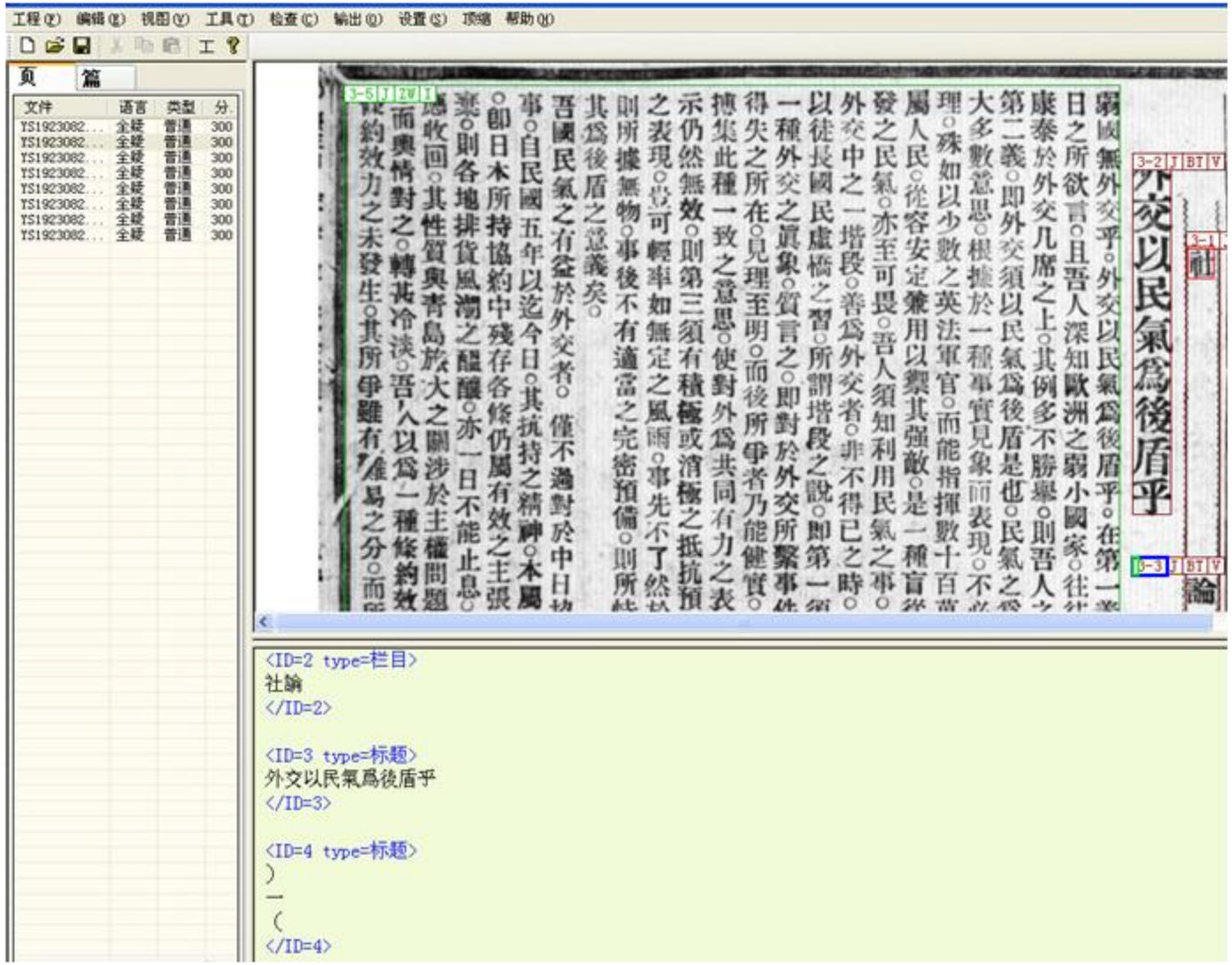
为能提高文字的质量，加大了字符聚类校对的比例，由普通繁体字项目的35%提高到了70%以上，保证文字的质量。

民国报纸数字资源建设

数字化加工

XML排版

为使成品XML结构正确，符合成品要求，针对文字进行格式排版

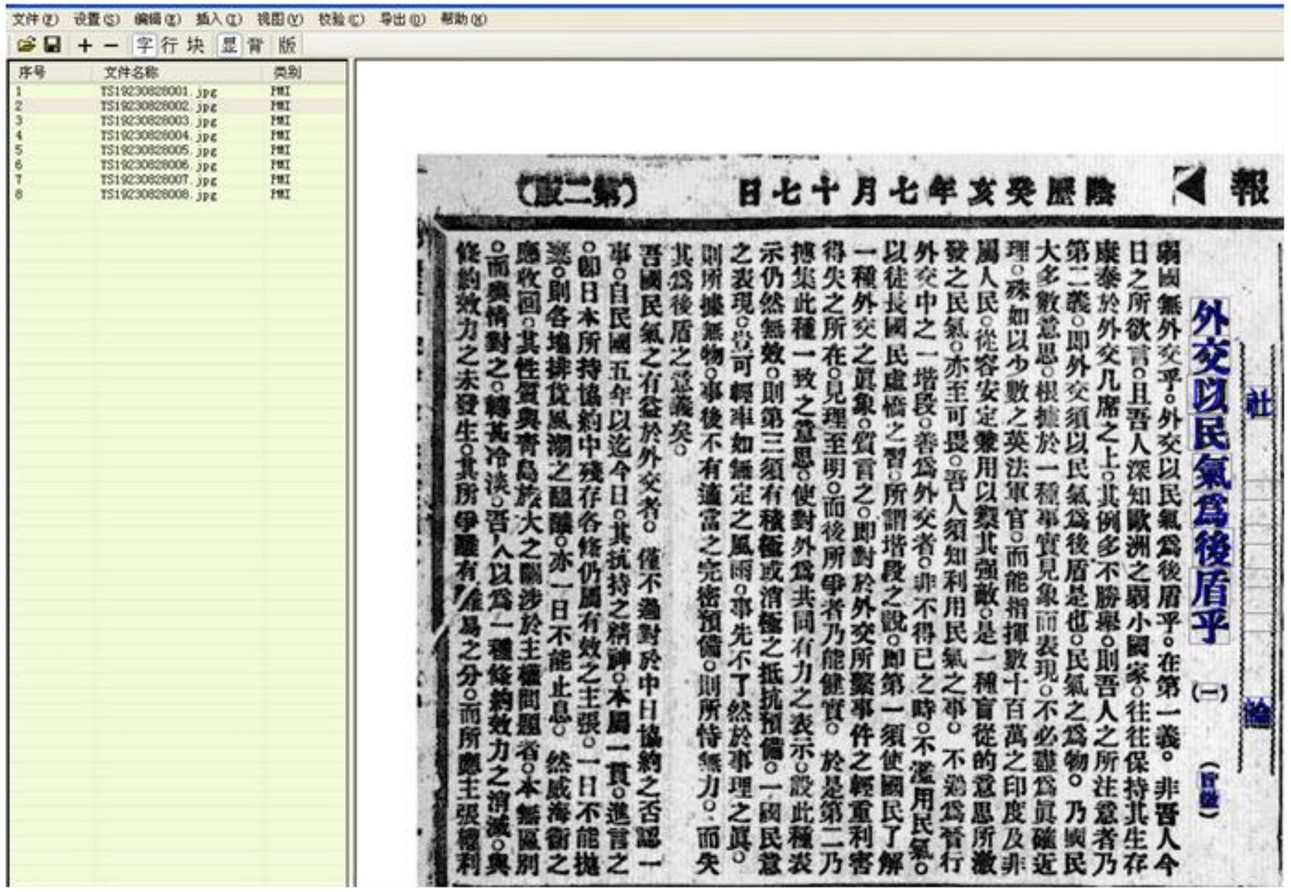


民国报纸数字资源建设

数字化加工

PDF排版

为使成品PDF对位准确，针对PDF进行了排版



XML成品

按照项目要求，导出符合要求的成品数据


```
<?xml version="1.0" encoding="utf-8" ?>
<root>
	<报纸元数据>
		<记录标识号>00N000203</记录标识号>
		<题名>益世报</题名>
		<出版日期>19230611</出版日期>
		<卷期>2598</卷期>
		<版次>7</版次>
		<整版PDF链接>007.pdf</整版PDF链接>
	</报纸元数据>
	<版元数据>
		<正文组>
			<正文 篇目号="001">
				<栏目>小言</栏目>
				<引题 />
				<标题>小學教員與軍警</标题>
				<副题 />
				<标题坐标>3121,437,3174,437,3174,794,3121,794</标题坐标>
				<篇目坐标>2180,334,3174,334,3174,1374,2180,1374</篇目坐标>
				<转版信息 />
				<作者 />
				<附图组 />
			</正文>
			<正文 篇目号="002">
				<栏目>京畿寫實</栏目>
				<引题 />
				<标题>殺夫案蒸骨驗屍</标题>
				<副题 />
				<标题坐标>1986,395,2036,395,2036,752,1986,752</标题坐标>
				<篇目坐标>404,286,2036,286,2036,1446,404,1446</篇目坐标>
				<转版信息 />
				<作者>正</作者>
				<附图组 />
			</正文>
			<正文 篇目号="003">
				<栏目>京畿寫實</栏目>
				<引题 />
				<标题>穆家園二次發現炸彈</标题>
				<副题 />
				<标题坐标>344,395,386,395,386,857,344,857</标题坐标>
				<篇目坐标>296,276,386,276,386,1430,296,1430;2627,1446,3326,1446,3326,2618,2627,2618</篇目坐标>
				<转版信息 />
				<作者>正</作者>
				<附图组 />
			</正文>
			<正文 篇目号="004">
				<栏目>京畿寫實</栏目>
				<引题 />
				<标题>爭家產演唱劈棺</标题>
				<副题 />
				<标题坐标>2554,1568,2616,1568,2616,1974,2554,1974</标题坐标>
				<篇目坐标>1931,1463,2616,1463,2616,2610,1931,2610</篇目坐标>
				<转版信息 />
				<作者>言</作者>
				<附图组 />
			</正文>
```

民国报纸数字资源建设

数字化加工

XML实例

```
<?xml version="1.0" encoding="utf-8"?>
<root>
  <报纸元数据>
    <记录标识号>00N001461</记录标识号>
    <题名>京报</题名>
    <出版日期>19190208</出版日期>
    <卷期>115</卷期>
    <版次>3</版次>
    <整版PDF链接>003.pdf</整版PDF链接>
  </报纸元数据>
  <版元数据>
    <正文组>
      <正文 篇目号="001">
        <栏目>要聞二</栏目>
        <引题 />
        <标题>京兆尹易人與張元奇</标题>
        <副题 />
        <标题坐标>2806,302,2855,302,2855,707,2806,707</标题坐标>
        <篇目坐标>2577,259,2860,259,2860,813,2577,813</篇目坐标>
        <作者 />
        <转版信息 />
        <附图组 />
      </正文>
      <正文 篇目号="002">
        <栏目>要聞二</栏目>
        <引题 />
        <标题>華工游歷英倫</标题>
        <副题 />
        <标题坐标>2522,302,2581,302,2581,583,2522,583</标题坐标>
        <篇目坐标>2206,258,2581,258,2581,807,2206,807</篇目坐标>
        <作者 />
        <转版信息 />
        <附图组 />
      </正文>
      <正文 篇目号="003">
        <栏目>要聞二</栏目>
        <引题 />
        <标题>審計院核銷之認眞</标题>
        <副题 />
        <标题坐标>2151,304,2205,304,2205,670,2151,670</标题坐标>
        <篇目坐标>1971,256,2210,256,2210,808,1971,808</篇目坐标>
        <作者 />
        <转版信息 />
        <附图组 />
      </正文>
```

成品质检

报纸提交之前，公司进行自检。

如右图所示，左侧是检查的数据列，中间是检查的成品内容（元数据内容与篇目内容），右侧展示的是每篇文章在整版中的坐标位置

XML质量要求

1、XML文件要求遵守XML语法规则，使用utf-8字符集

2、XML文件应著录全面的报纸信息，记录标题及篇目位置等信息，标题位置为引题、标题和副题的整体位置，遵守XML语法规则，按规范标签进行制作

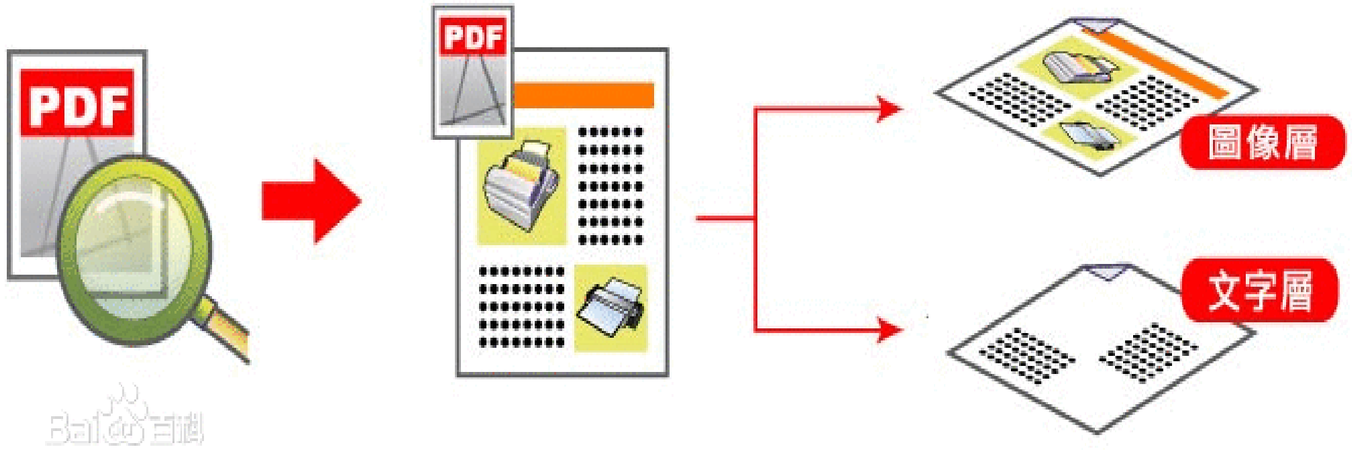
3、文件命名无误，且在数量上与TIFF图像一致。

4、XML数据内容与TIFF图像内容吻合，不存在乱码、转换错误等问题。

5、XML数据应如实反映原版的篇信息，不应出现与篇目不符的字符、段落、硬回车、空格等。

双层PDF

只需要对篇目标题、
作者、栏目进行识别。



PDF质量要求

1、双层PDF的上层图像为JPG格式，下层为篇目标题文本层，且标题位置已置标。

2、PDF格式，要求PDF编码为1.5版本（兼容Adobe reader6.0），要求在保持图像清晰可读的基础上尽可能减小存储量。

3、PDF文字层所使用的字体以“已嵌入子集”方式嵌入PDF文件

4、双层PDF文件的图像层和文字层的文字对位准确，反显区域与文字区域相差1毫米以内。

5、双层PDF错误率不超过0.3‰。

民国报纸数字资源建设

数字化加工

PDF常见问题

图像层

文字层

右第一圖某工人死後臉已膨脹其胸間脅下起顆粒之白點乃被火燒灸所致口開作哭狀極爲可怖

（唐山煤礦工人慘狀一）

（未完）

⊙統一聲中之西南態度

之機關仍未廢止

和會正式公開之和會爲得當

軍府依然存在

正式宣告

此次北方宣言文等不能承認（下略）

莫榮新

本月敬日起率同將士宣

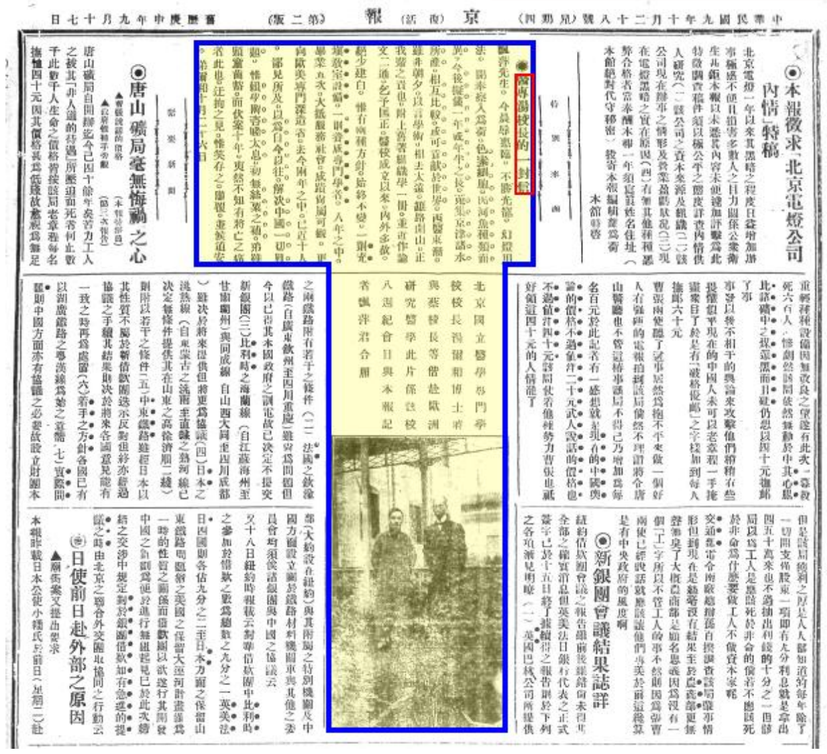
民国报纸数字资源建设

数字化加工

特殊情况的处理

- 附图与图片组

```
</正文组>
<图片组>
  <图片 篇目号="019">
    <图片标题>[新聞圖片]</图片标题>
    <标题坐标 />
    <篇目坐标>1503,1002,1503,2535,890,2535,890,1002</篇目坐标>
    <图片作者 />
  </图片>
</图片组>
```

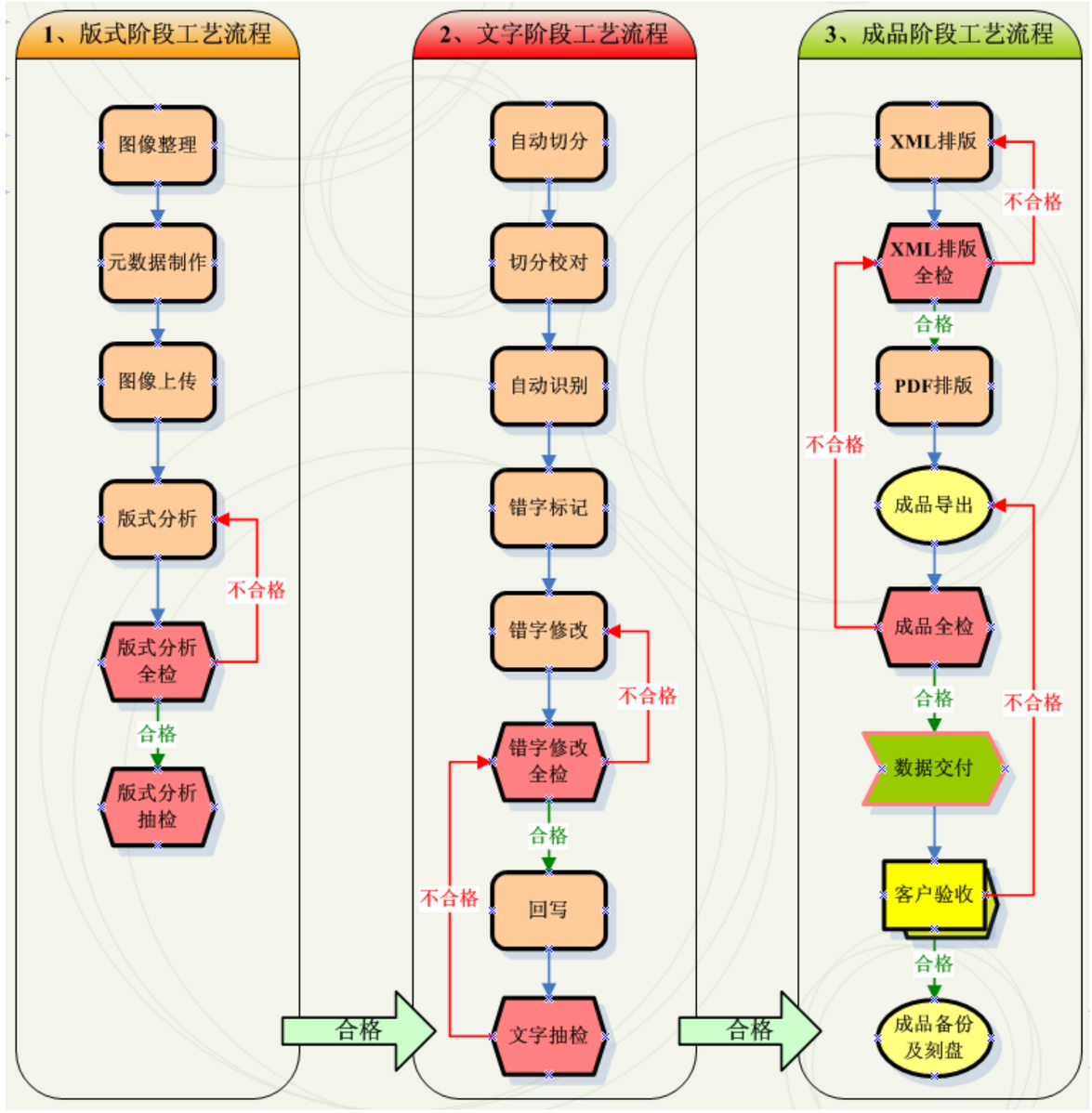


```
  <附图组 />
</正文>
<正文 篇目号="002">
  <栏目>特別来函</栏目>
  <引题 />
  <标题>醫專湯校長的一封信</标题>
  <副题 />
  <标题坐标>1849,302,1909,302,1909,722,1849,722</标题坐标>
  <篇目坐标>727,191,1909,191,1909,970,1813,970,1813,2610,1190,2610,1190,970,727,970</篇目坐标>
  <作者 />
  <转版信息 />
  <附图组>
    <附图 附图号="002_01" 版次="2">
      <附图坐标>1813,1027,1813,2610,1190,2610,1190,1027</附图坐标>
    </附图>
  </附图组>
</正文>
```

工艺流程

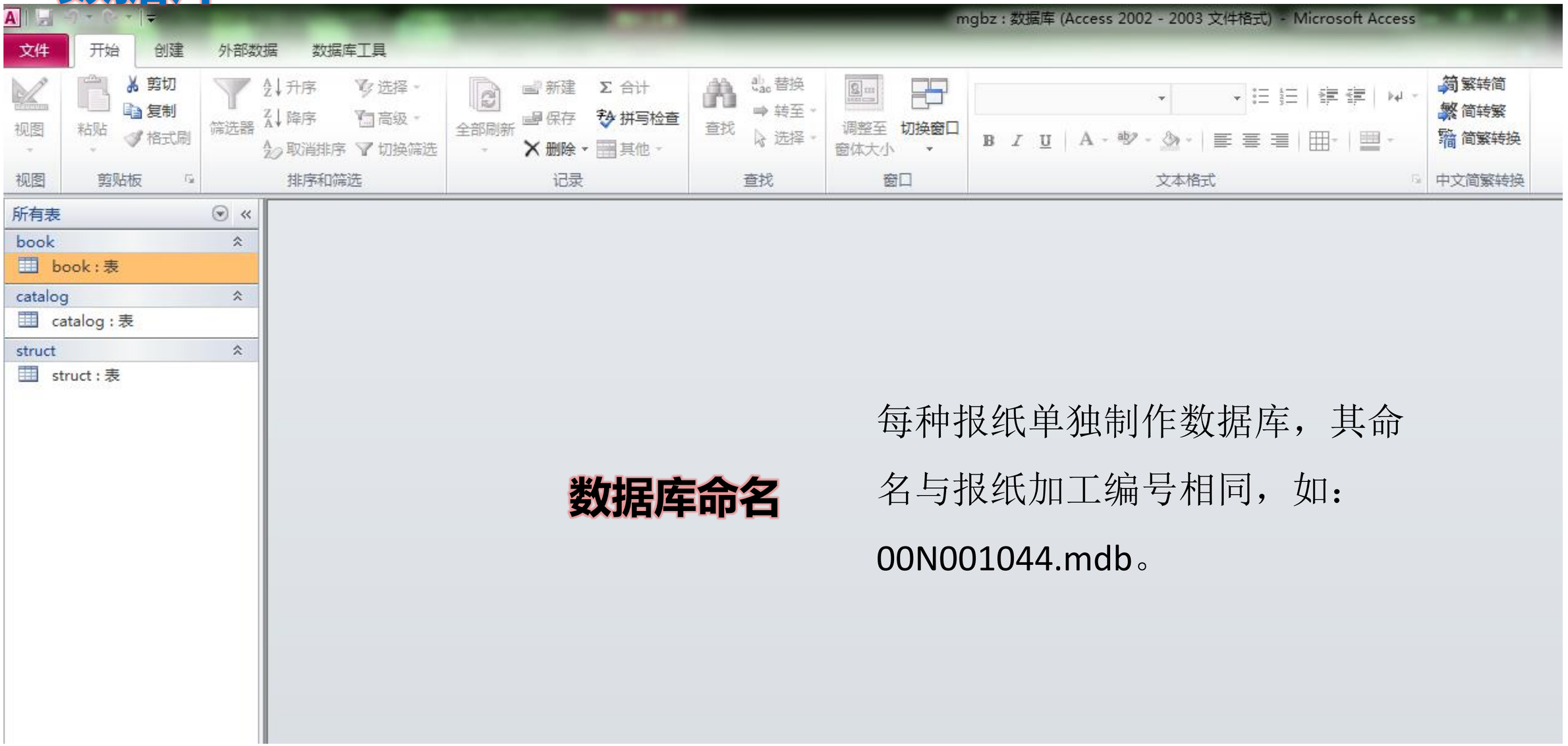
根据地方报纸的特点，通过合理的工艺流程来规避质量风险。

工艺流程如右图所示：



相关文档

数据库



数据库命名

每种报纸单独制作数据库，其命名与报纸加工编号相同，如：00N001044.mdb。

民国报纸数字资源建设

数字化加工

数据库

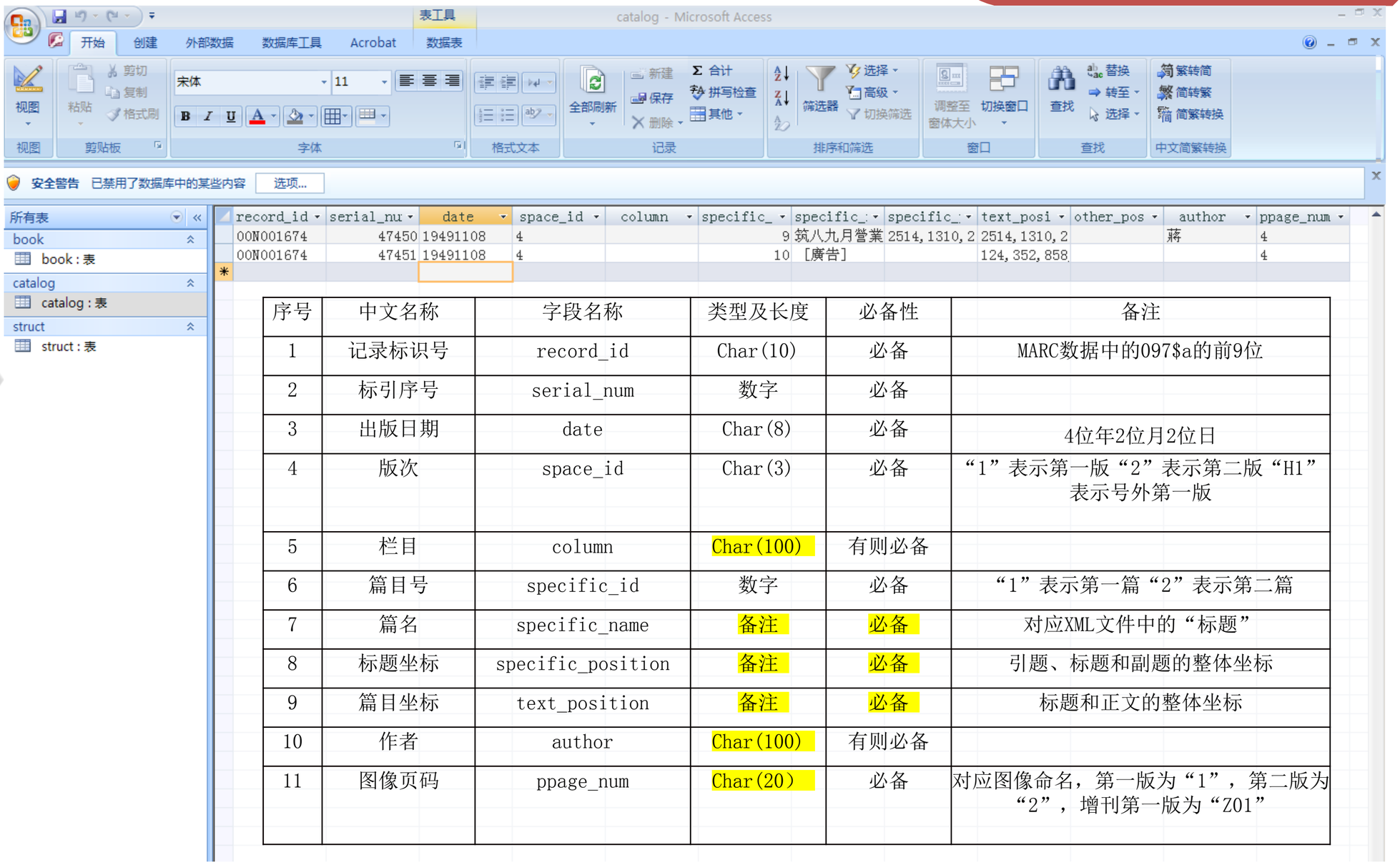
book表

序号	中文名称	字段名称	类型及长度	必备性	对应书目数据MARC内容
1	001	001	Char(10)	必备	MARC数据中的001字段
2	记录标识号	record_id	Char(9)	必备	MARC数据中的097$a的前9位
3	题名	book_name	Char(100)	必备	MARC数据中的200$a
4	出版地点	pub_place	Char(60)	有则必备	MARC数据中的210$a
5	出版者	pub_house	Char(60)	有则必备	MARC数据中的210$c
6	出版时间	pub_date	Char(60)	有则必备	MARC数据中的210$d
7	出版周期	pub_T	Char(60)	有则必备	MARC数据中的326$a $b（著录形式为a(b)；326字段若重复，用空格隔开）
8	起止卷期	start_end_day	备注	有则必备	ARC数据中的207$a（英文分号分隔$a重复字段）
9	更名信息	name_change	Char(100)	有则必备	MARC数据中的311$a
10	附注信息（号外、增刊等）	general_notes	Char(100)	有则必备	MARC数据中的300$a
11	特殊附注（出版规律）	specific_notes	备注	有则必备	MARC数据中的315$a
12	总卷期	pub_num	Char(60)	有则必备	MARC数据中的215$a
13	主题词	topic_word	Char(60)	有则必备	MARC数据中的610$a

民国报纸数字资源建设

数字化加工

数据库



catalog表

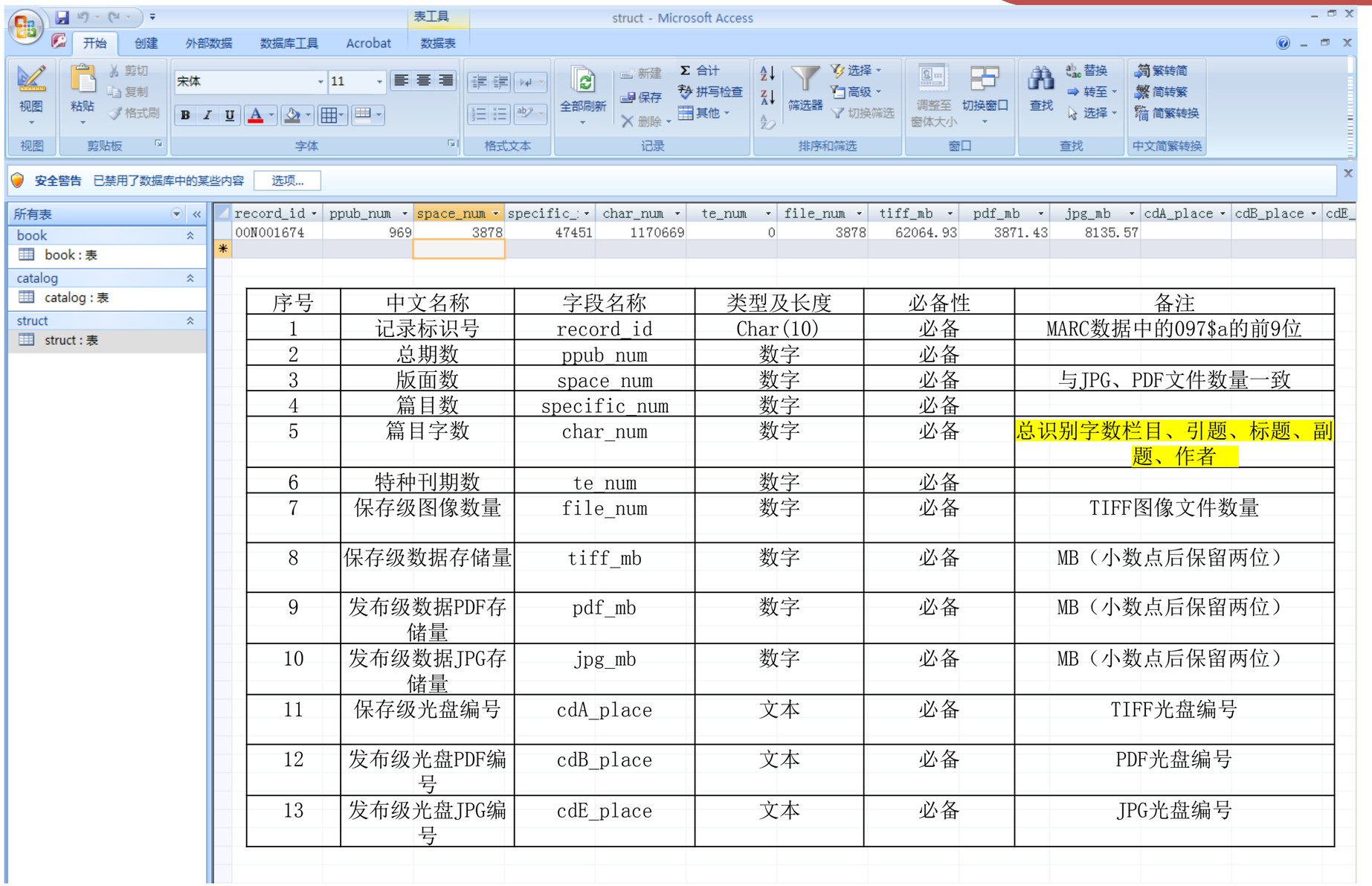
序号	中文名称	字段名称	类型及长度	必备性	备注
1	记录标识号	record_id	Char(10)	必备	MARC数据中的097$a的前9位
2	标引序号	serial_num	数字	必备	
3	出版日期	date	Char(8)	必备	4位年2位月2位日
4	版次	space_id	Char(3)	必备	“1”表示第一版“2”表示第二版“H1”表示号外第一版
5	栏目	column	Char(100)	有则必备	
6	篇目号	specific_id	数字	必备	“1”表示第一篇“2”表示第二篇
7	篇名	specific_name	备注	必备	对应XML文件中的“标题”
8	标题坐标	specific_position	备注	必备	引题、标题和副题的整体坐标
9	篇目坐标	text_position	备注	必备	标题和正文的整体坐标
10	作者	author	Char(100)	有则必备	
11	图像页码	ppage_num	Char(20)	必备	对应图像命名，第一版为“1”，第二版为“2”，增刊第一版为“Z01”

民国报纸数字资源建设

数字化加工

数据库

struct表



序号	中文名称	字段名称	类型及长度	必备性	备注
1	记录标识号	record_id	Char(10)	必备	MARC数据中的097$a的前9位
2	总期数	ppub_num	数字	必备	
3	版面数	space_num	数字	必备	与JPG、PDF文件数量一致
4	篇目数	specific_num	数字	必备	
5	篇目字数	char_num	数字	必备	总识别字数栏目、引题、标题、副题、作者
6	特种刊期数	te_num	数字	必备	
7	保存级图像数量	file_num	数字	必备	TIFF图像文件数量
8	保存级数据存储量	tiff_mb	数字	必备	MB（小数点后保留两位）
9	发布级数据PDF存储量	pdf_mb	数字	必备	MB（小数点后保留两位）
10	发布级数据JPG存储量	jpg_mb	数字	必备	MB（小数点后保留两位）
11	保存级光盘编号	cdA_place	文本	必备	TIFF光盘编号
12	发布级光盘PDF编号	cdB_place	文本	必备	PDF光盘编号
13	发布级光盘JPG编号	cdE_place	文本	必备	JPG光盘编号

数据库制作要求

1、标引数据库以MDB数据库方式提交，后缀名mdb；

2、基本信息表的内容应与MARC数据相应内容保持一致；

3、版面篇目信息数据库标引要求真实反映报纸原貌；

4、结构信息表应严格按文献实际情况进行描述；

5、无法录入的生僻字等用“〓”表示；

6、版面篇目信息表与XML文件的对应元素项的内容应一致；

7、各种著录、说明文件的文字、符号、版式、位置和文件名称准确，

8、其综合错误率不超过0.3‰。

说明文件

文献总体说明文件

数据总体说明
文献单册数据量统计
保存级对象数据硬盘存储清单
发布级对象数据硬盘存储清单

单册文献说明文件

报纸的名称、总期数和版面数等

存储介质说明文件

存储介质信息：文献数量、文件数量、存储容量等；

技术参数：存储格式、加工设备、加工软件、扫描方式、扫描分辨率等

民国报纸数字资源建设

- 文献前整理
- 数字化加工
- 数据验收

数据验收

- 规范性
- 一致性
- 正确性
- 完整性
- 有效性

数据验收

- 依据图像质量，一般采用通查和抽查相结合的方式进行地方报纸数据验收。

基本内容

- 验收方式；
- 验收比例；
- 验收内容；
- 辅助工具；
- 验收报告。

提交格式

```
├─21160000、readme.txt
|    ├──TIFF
|    |      ├─加工编号
|    |      |     ├─日期
|    |      |     |     ├001.tif
|    |      |     |     ├002.tif……
|    |      |     ├─日期……
|    |      ├─加工编号
|    |      |     ├─日期
|    |      |     ├─日期……
|    |      └─……
|    ├──PDF
|    |      ├─加工编号
|    |      └─加工编号……
|    ├──XML
|    |      ├─加工编号
|    |      └─加工编号……
|    ├──21160000.iso
|    ├──21160000_001.mdb
|    └──21160000.xls
```

民国报纸数字资源建设

数据验收

对象数据

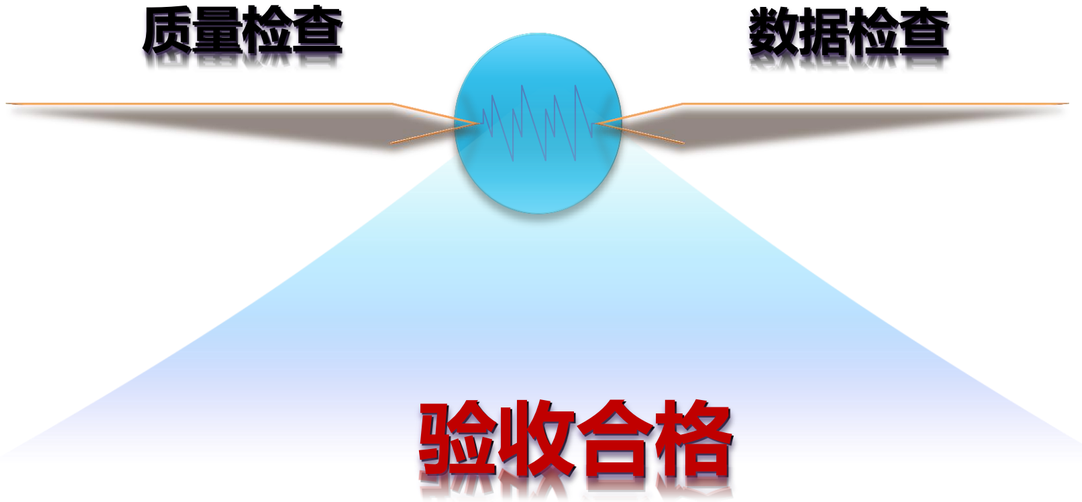
- tiff 图像
- jpg 图像
- 双层 pdf 文件
- Xml 文件

数据库及相关说明文件

- 对应数据库
- 介质说明文件
- 民国报纸数据说明
- 单册文献说明文件
- 报纸书目数据

数据提交相关文件

- 第三方质检报告
- 民国报纸验收数据提交单



错误实例

原件问题

序号	包名	问题
1	大刚报19460125	原1月25日第4版印刷成1月24日
2	大刚报19460202	第1版图像有2张，选用最佳图像
3	大刚报19460323	原3月23日第3版印刷成3月26日
4	大刚报19460520	第2版版次印刷成第3版
5	大刚报19460609	第2版没有印刷日期，第3版印刷成第2版）
6	大刚报19460610	第3版版次印刷成第2版
7	大刚报19460627	原图像残缺，看不见版次日期
8	大刚报19460706	第3版版次印刷成第2版
9	大刚报19460723	第3版没有印刷版次
10	大刚报19460809	原图像残缺，第3版版次看不清
11	大刚报19460815	第3版没有印刷版次
12	大刚报19460817	第5版版次印刷成第6版
13	大刚报19460902	原9月2日第4版印刷成9月1日，第6版印刷成第3版
14	大刚报19460903	第2版版次印刷成第6版
15	大刚报19460912	原9月12日第1版印刷成9月13日
16	大刚报19461015	缺少1、2、3、4版，只有5、6版
17	大刚报19461019	原10月19日第4版印刷成10月18日第2版
18	大刚报19461020	第4版版次印刷成第5版

民国报纸数字资源建设

数据验收

通查内容

result.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

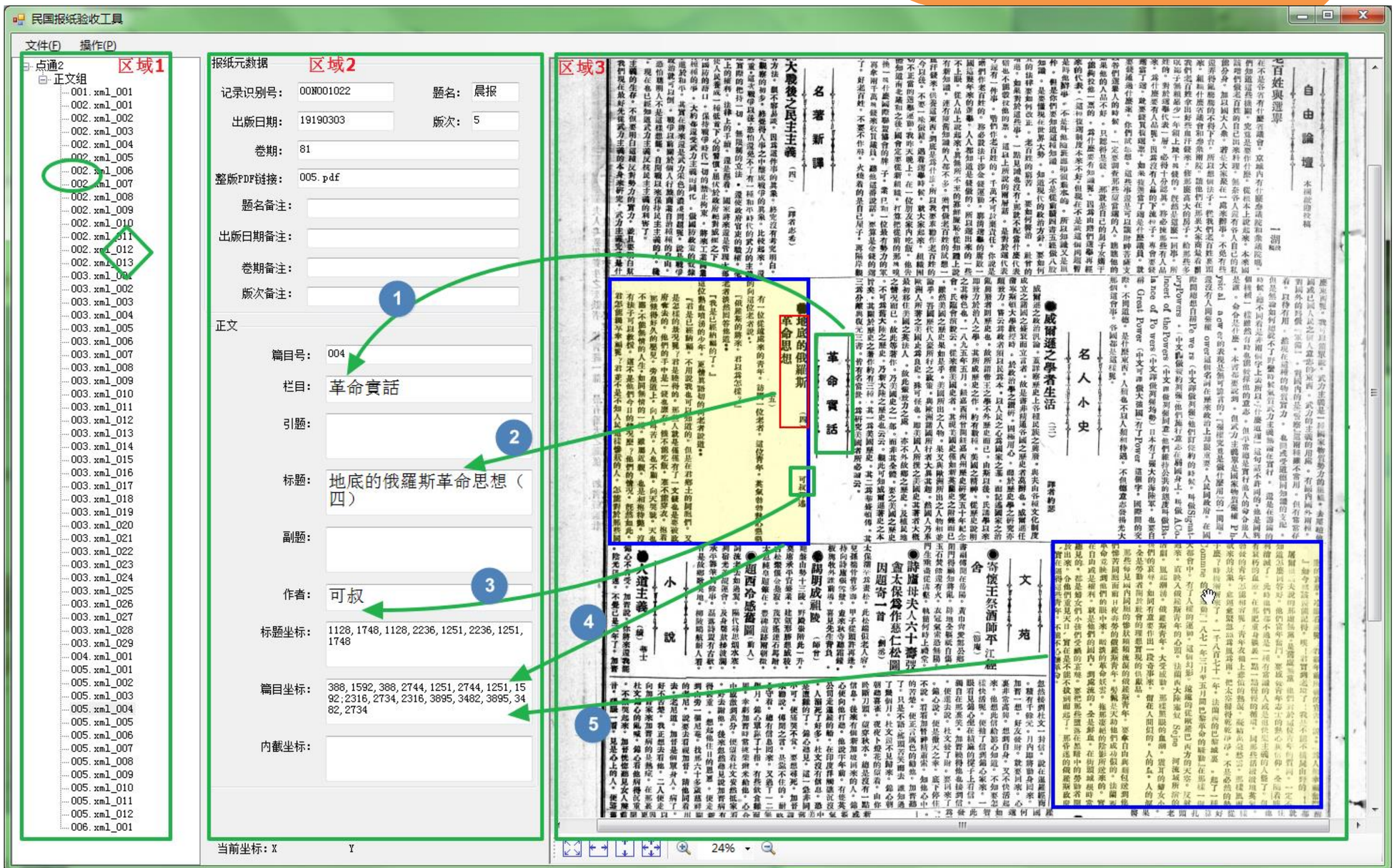
```
H:\晨报\xml\00N001022

**********19181201**********

002.xml第2个正文的第1个附图的“附图坐标”  :  不存在
003.xml第9个正文的第1个附图的“附图坐标”  :  不存在

**********19181219**********

003.xml第1个正文的第1个附图的“附图坐标”  :  不存在
**********19181227**********

007.xml第2个正文的第1个附图的“附图坐标”  :  不存在
**********19190105**********

003.xml文件                :  加载失败
007.xml文件                :  加载失败
**********19190106**********

002.xml文件                :  加载失败
003.xml文件                :  加载失败
005.xml文件                :  加载失败
```

ID	文件路径	图像文件名	图像扩展名	扩展名大小	图像类型	文件大小(M	图像宽度	图像高度	像素总数	位深度	压缩方式	页面个数	水平分辨率	垂直分辨率
7960	I:\第5批成品	001.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7961	I:\第5批成品	002.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7962	I:\第5批成品	003.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7963	I:\第5批成品	004.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7964	I:\第5批成品	F01.tif	tif	小写	TIFF	.41	2506	3529	8843674	1	LZW	1	300	300
7965	I:\第5批成品	F02.tif	tif	小写	TIFF	.39	2506	3529	8843674	1	LZW	1	300	300
7966	I:\第5批成品	001.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7967	I:\第5批成品	002.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7968	I:\第5批成品	003.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7969	I:\第5批成品	004.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7970	I:\第5批成品	001.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7971	I:\第5批成品	002.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7972	I:\第5批成品	003.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7973	I:\第5批成品	004.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7974	I:\第5批成品	001.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7975	I:\第5批成品	002.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7976	I:\第5批成品	003.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7977	I:\第5批成品	004.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7978	I:\第5批成品	001.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7979	I:\第5批成品	002.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300
7980	I:\第5批成品	003.tif	tif	小写	TIFF	16.89	5013	3529	17690877	8	None	1	300	300

民国报纸数字资源建设

数据验收

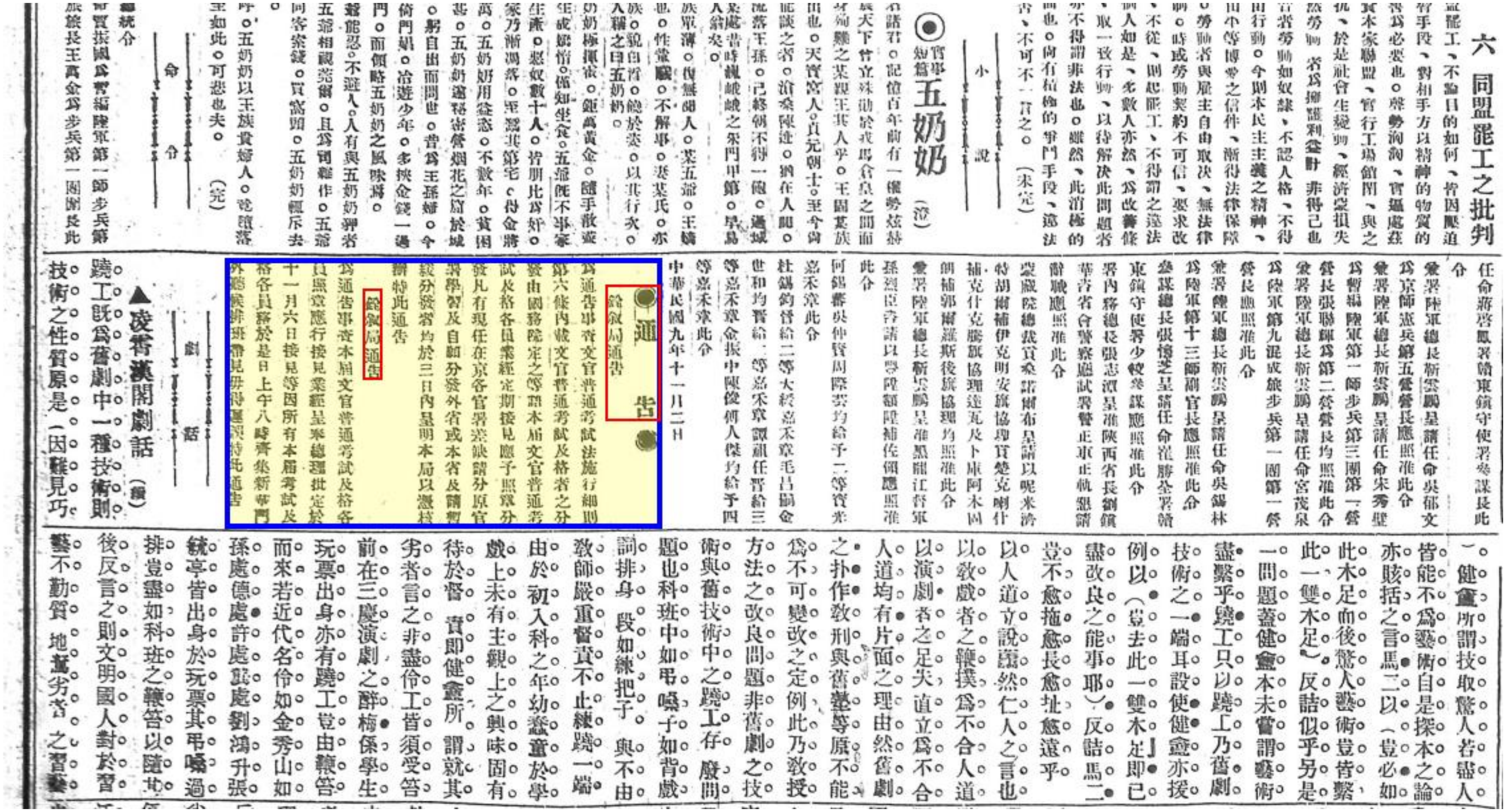
借助工具



民国报纸数字资源建设

数据验收

栏目错误



六 同盟罷工之批判

小說

五奶奶

通告

銓敘局通告

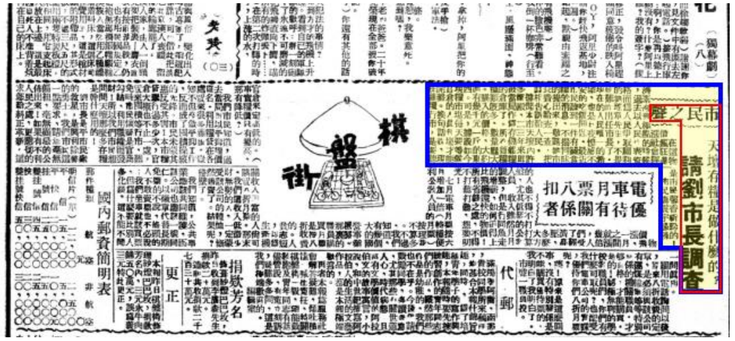
劇話

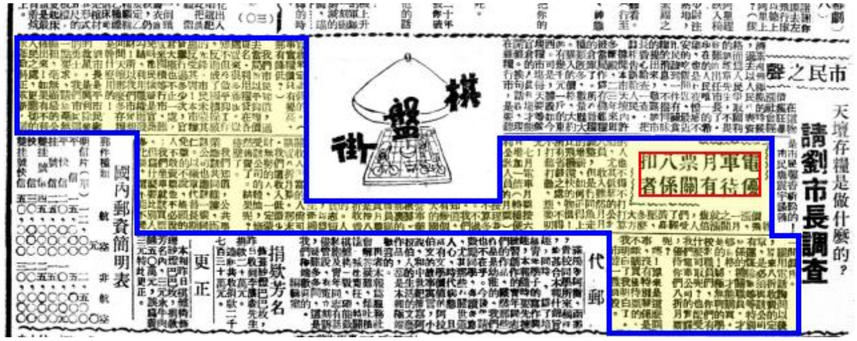
廢書齋閑劇話

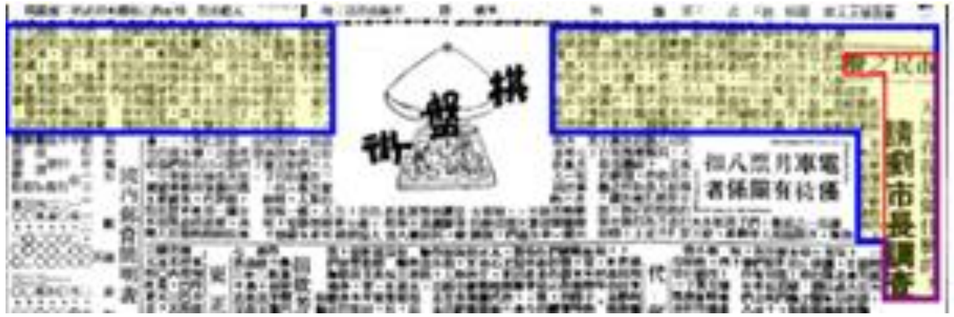
民国报纸数字资源建设

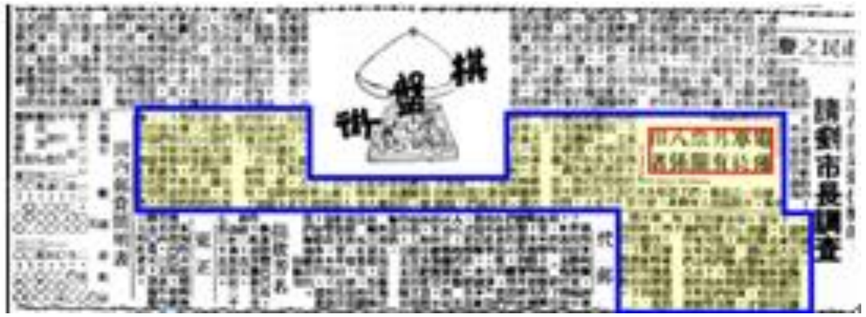
数据验收

篇目错误





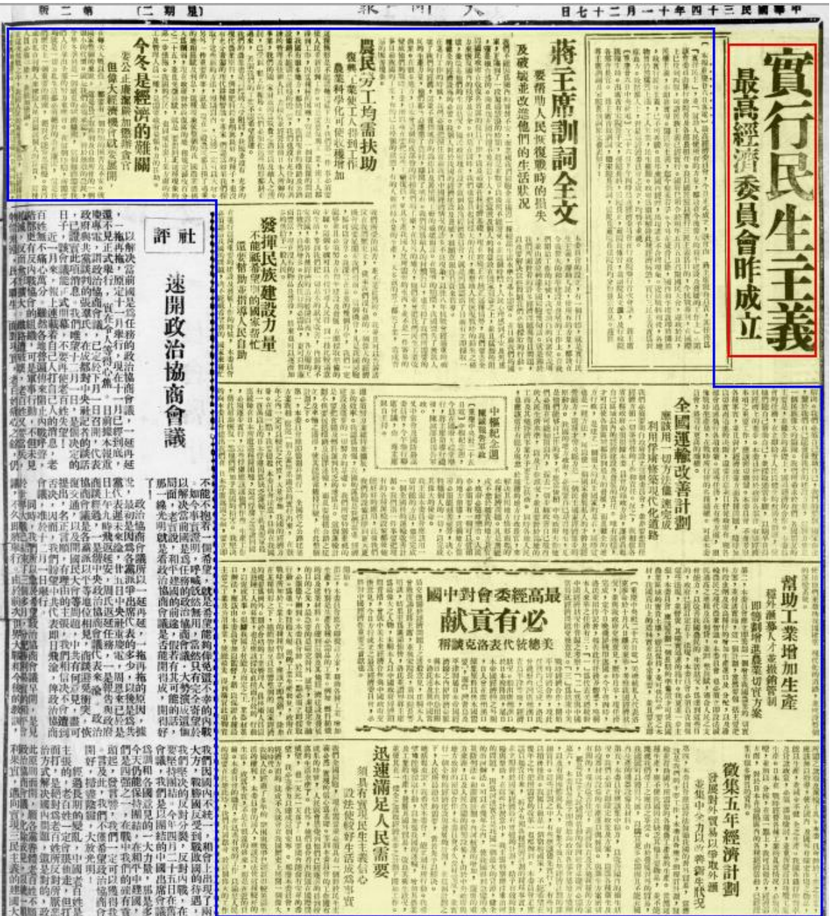




民国报纸数字资源建设

数据验收

内截错误


實行民生主義
最高經濟委員會昨成立
蔣主席訓詞全文
今冬是經濟的難關
社評
速開政治協商會議
農民分工均需扶助
發揮民族建設力量
全國運輸改善計劃
迅速滿足人民需要
徵集五年經濟計劃


國中對會委經高最
必有貢獻
稱謂克洛表代統總美

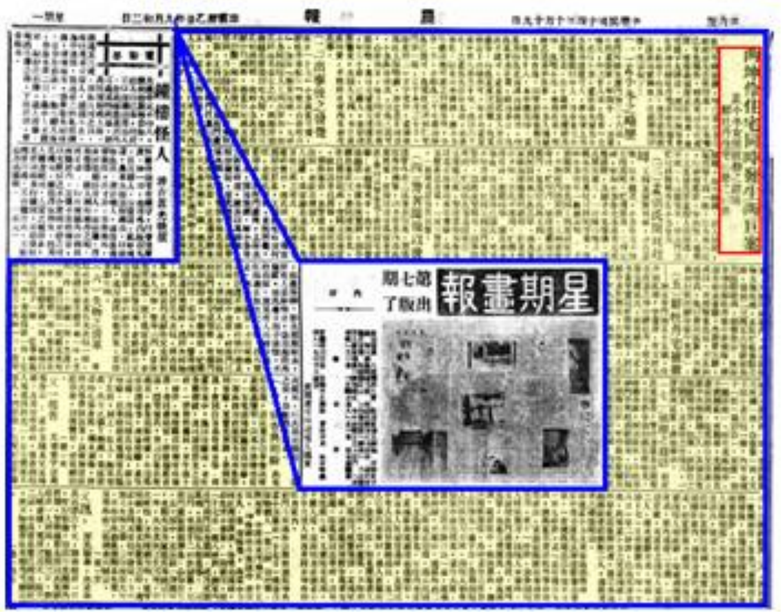

全國運輸改善計劃
國中對會委經高最
必有貢獻
稱謂克洛表代統總美

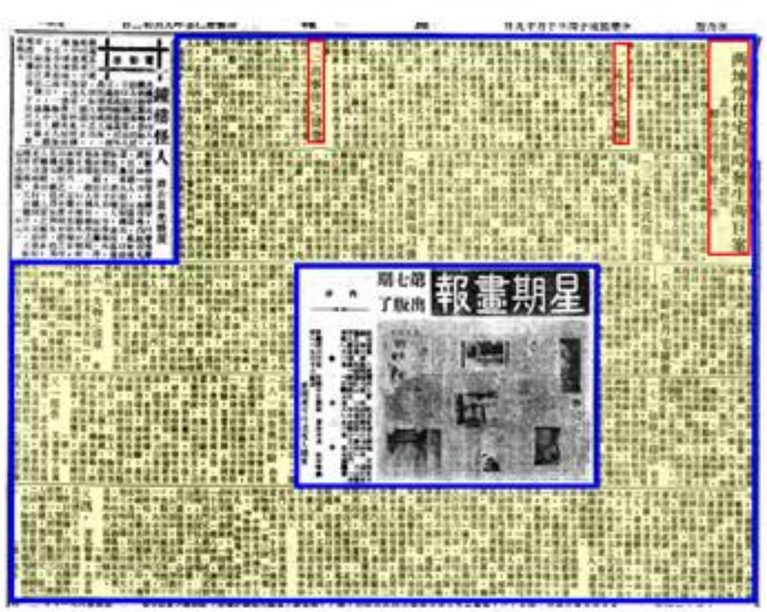

實行民生主義

民国报纸数字资源建设

数据验收

置标混乱





民国报纸数字资源建设

数据验收

置标混乱

特寫

周曼華北來已不可能

上海之銀色動態一斑

影星北來成謎局

說戲

華北運動大會

競賽程序公佈

各項裁判員昨亦聘妥

大會定七日開始舉行

國立院校

市屬各校

興亞展覽延期一週

民国报纸数字资源建设

数据验收

标题不完整

上訴三審案件之處理

司法委員會 議定辦法 董康委員長在常會提出

宣言

絕不承認

鄭州市民紛紛避難

上海外匯再跌

廈門稅關

開封市面恢復

各雜軍不滿

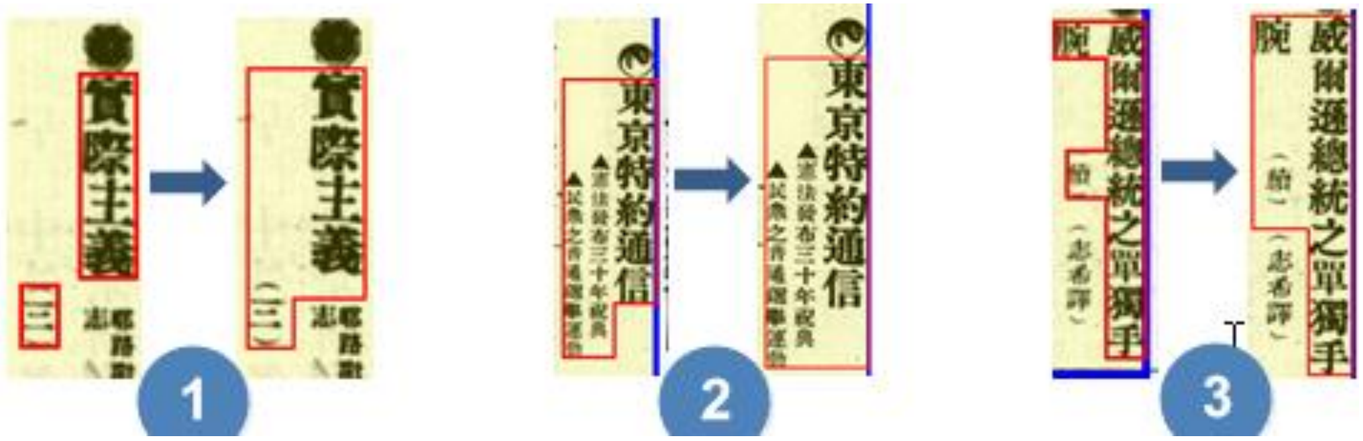
七省截用鹽款之細數

秦閩川粵滇湘綏遠等七處提支鹽稅表

七省截用鹽款之細數

秦閩川粵滇湘綏遠等七處提支鹽稅表

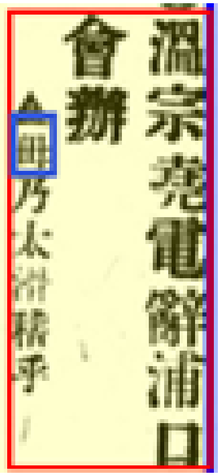
置标不美观



数据验收

识别错误

副題： 各地散發傳單「打倒德國」貴族及軍界人士多告失蹤



副題： 母乃太滑稽乎

标题： 大連會議俄代表是出最後答案


谢谢！